Conformal Predictions in Multimedia Pattern Recognition

by

Vineeth Nallure Balasubramanian

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

ARIZONA STATE UNIVERSITY

December 2010

Conformal Predictions in Multimedia Pattern Recognition

by

Vineeth Nallure Balasubramanian

has been approved

September 2010

Graduate Supervisory Committee:

Sethuraman Panchanathan, Chair
Jieping Ye
Baoxin Li
Vladimir Vovk

ACCEPTED BY THE GRADUATE COLLEGE

ABSTRACT

The fields of pattern recognition and machine learning are on a fundamental quest to design systems that can learn the way humans do. One important aspect of human intelligence that has so far not been given sufficient attention is the capability of humans to express when they are certain about a decision, or when they are not. Machine learning techniques today are not yet fully equipped to be trusted with this critical task. This work seeks to address this fundamental knowledge gap. Existing approaches that provide a measure of confidence on a prediction such as learning algorithms based on the Bayesian theory or the Probably Approximately Correct theory require strong assumptions or often produce results that are not practical or reliable. The recently developed Conformal Predictions (CP) framework - which is based on the principles of hypothesis testing, transductive inference and algorithmic randomness - provides a game-theoretic approach to the estimation of confidence with several desirable properties such as online calibration and generalizability to all classification and regression methods.

This dissertation builds on the CP theory to compute reliable confidence measures that aid decision-making in real-world problems through: (i) Development of a methodology for learning a kernel function (or distance metric) for optimal and accurate conformal predictors; (ii) Validation of the calibration properties of the CP framework when applied to multi-classifier (or multi-regressor) fusion; and (iii) Development of a methodology to extend the CP framework to continuous learning, by using the framework for online active learning. These contributions are validated on four real-world problems from the domains of healthcare and assistive technologies: two classification-based applications (risk prediction in cardiac decision support and multimodal person recognition), and two regression-based applications (head pose estimation and saliency prediction in images). The results obtained show that: (i) multiple kernel learning can effectively increase efficiency in the CP

framework; (ii) quantile p-value combination methods provide a viable solution for fusion in the CP framework; and (iii) eigendecomposition of p-value difference matrices can serve as effective measures for online active learning; demonstrating promise and potential in using these contributions in multimedia pattern recognition problems in real-world settings.

ACKNOWLEDGEMENTS

Over the last few years, my PhD dissertation has provided me with wonderful opportunities to be mentored by, to interact with, and to be supported by some of the finest minds and personalities that I have come across in my life. I would like to take this opportunity to thank every one of them with all my heart.

This work would never have been possible without the generous guidance and support of my mentor and advisor, Prof. Sethuraman Panchanathan, who magnanimously gave me the freedom to pursue my research interests (and let me 'feel free like a bird', in his words). I cannot thank him enough for his strong belief in me over the years, for setting standards of excellence that will take me a lifetime to scale, for housing me in an environment suffused with bright minds and numerous opportunities for exposure and growth, and most importantly, for his never-failing support through every high and low of my PhD.

I would like to thank my committee members, Dr Jieping Ye, Dr Baoxin Li and Dr Vladimir Vovk, for their kindness in sparing their valuable time to interact with me whenever I needed, and for sharing inputs that have shaped my thinking - not only from an academic perspective, but also all-round development. Throughout my PhD years, I have always looked forward to interacting with each one of them and I consider it my privilege to have worked with them. My special thanks to Dr Vovk, who agreed to serve on my committee despite the geographical distance, and provided valuable inputs that made this dissertation come alive.

It has been a great pleasure working with fellow members of the Center for Cognitive Ubiquitous Computing (CUbiC) at Arizona State University. I would like to convey my sincere gratitude to Shayok for having supported me with my research at every stage from inception to completion; to Sreekar, CK and Troy for all those memorable moments of working together on proposals and write-ups; to Ramkiran and Sunaad for bearing with me all through their theses; to Mohammad,

David, Mike, Karla, Rita, Daniel, Ashok, Prasanth, Hiranmayi, Jeff, Jessie and Cindy, for all their help, insights and most of all, cheer. I would also like to thank all the faculty, staff, and students at Arizona State University for providing me with all the necessary support during my tenure as a doctoral student.

My research has benefited tremendously from various collaborations over the years. I would particularly like to thank Dr. Ambika Bhaskaran, Jennifer Vermillion, Jenni Harris (at Advanced Cardiac Specialists); Prof. Juan Nolazco, Paola Garcia, Roberto Aceves (at Tecnologico de Monterey, Mexico); Dr John Black, Dr Terri Hedgpeth, Dr Dirk Colbry, Dr Gaurav Pradhan (at CUbiC); Dr Calliss, Prof Nielsen, Dr Konjevod (Computer Science department, ASU) for the many hours of thoughtful conversations. In particular, I would like to thank John and Terri for their selfless guidance and support during my initial years, when their kindness and concern truly made CUbiC a second home.

My dissertation research has been sponsored by grants from National Science Foundation (NSF-ITR grant IIS-0326544 and NSF IIS-0739744) and the Arizona State University Strategic Investment Fund. I sincerely thank the NSF and the ASU Office of Knowledge Enterprise Development for their kind support.

My heartfelt thanks is due to all my friends and acquaintances in India and USA, who have suffused my life with their warmth and concern. My deep gratitude to CK, Shreyas and Ramkiran - my roommates during my initial PhD years, who enriched my life in many different ways, and left me with wonderful memories of good times.

Lastly, but most importantly, I would not be what I am today without the support, care and love that I receive from my family. To Padmini, Siki and Vidya, words fail to express my gratitude. To my dear parents and Swami, although this may be an imperfect piece of work, I dedicate this to you.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Chapter 1

INTRODUCTION AND MOTIVATION

Over the centuries of human existence, the recognition of patterns in observed data has led to numerous discoveries, and has eventually paved the path to the development of vast bodies of scientific knowledge. As pointed out by Bishop [10], the study of observational data has led to the discovery of various phenomena in fields ranging from astronomy to avian life to atomic spectra, including the understanding of the laws of planetary motion, migratory patterns of birds and the development of quantum physics. However, the field of pattern recognition has relied immensely on manual expertise and experience over the bygone centuries. With the tremendous growth of computing resources and algorithms, the last 50 years have re-defined the field of pattern recognition as the automatic discovery of patterns in observed data through the use of computer algorithms.

Over the last few decades, multimedia computing has experienced an explosive growth in terms of generation of data in various modalities such as text, images, video, audio and now, haptics (the sense of touch). This has led to the extensive use of pattern recognition techniques in multimedia computing, but the rate of generation of multimedia data has sustained an equivalent increasing need for intelligent computer algorithms that can automatically identify regularities in data - thereby creating newer challenges that need to be addressed by researchers in pattern recognition.

The success of automatic pattern recognition in recent decades has relied on the use of machine learning techniques to automatically learn to categorize data. Machine learning aims at the design and development of algorithms that automatically

learn to recognize complex patterns and make intelligent decisions based on data. Machine learning approaches have led to numerous successes in pattern recognition in varied applications such as digit recognition, spam filtering, face detection, fault detection in industrial manufacturing, and many others [11]. However, complex real-world problems (such as face recognition or patient risk prognosis) are associated with several factors causing uncertainty in the decision-making process, and assumptions are often made to resolve the uncertainty. In order to help end users with decision-making in such complex problems, it has become very essential to compute a *reliable* measure of confidence that expresses the belief of the algorithm in the predicted result. By this measure is meant a unique single numeric value ($\in [0, 1]$) that is associated with a prediction on a given test data point, and provide a measure of belief of the learning system on a hypothesis, given the evidence, as defined by Cheeseman [12]. While earlier work in related areas use different, yet closely associated, terms such as '*belief*' or '*reliability*', the term '*confidence*' is used in this work, and for this purpose, considered synonymous to belief or reliability. The design and development of efficient algorithms for multimedia pattern recognition that can compute a reliable measure of confidence on their predictions is the underlying motivation of this work.

1.1 Uncertainty Estimation: An Overview

The estimation of uncertainty has been extensively studied from different perspectives for over half a century now. The application of computational methods in fields ranging from seismology to finance has made uncertainty quantification a universally relevant topic. Existing literature in uncertainty quantification segregates uncertainty into two main kinds, as listed in Table 1.1. The approaches typically used to address each of these kinds of uncertainty are also mentioned in Table

1.1. While *aleatory uncertainty* is difficult to resolve, most existing approaches in related fields attempt to address *epistemic uncertainty*. A detailed review of these sources is presented by Daneshkhah in [13].

Type of uncertainty	Description	Approaches used
Aleatory/ Statistical	Arises due to natural, unpredictable variations in the system under study. Also called irreducible uncertainty.	Techniques such as Monte Carlo simulation are used to capture statistical variations. Probability density functions such as Gaussian are often represented by their moments (such as mean and variance). More recently, Karhunen-Loeve and polynomial chaos expansions are used for this purpose.
Epistemic/ Systematic	Arises due to a lack of knowledge about the behavior of the system, and can be conceptually resolved.	Methods such as fuzzy logic or evidence theory are used to resolve such uncertainty.

Table 1.1: Types of uncertainty

Given these basic categories of uncertainty, we now present an overview of the sources of uncertainty, the approaches to uncertainty estimation and the representations of uncertainty (as commonly used in pattern recognition and machine learning) in the following subsections.

Sources of Uncertainty

Uncertainty, in the context of multimedia pattern recognition, arises from many sources, such as: (i) the inherent limitations in our ability to model the world, (ii) noise and perceptual limitations in sensor measurements, or (iii) the approximate nature of algorithmic solutions [14]. With respect to traditional pattern recognition and machine learning approaches, these sources of uncertainty can be categorized

in the following manner (a similar categorization is also presented by Shrestha and Solomatine in [15]):

- **Data Uncertainty:** Often, the data used in applications is a significant source of uncertainty. Data may be noisy, may have missing values, may contain anomalies (such as a particular data value exceeding the range suggested for the attribute), or may contain attributes that are highly correlated (while the algorithm assumes independence of the attributes).
- **Model Uncertainty:** The model structure, i.e., how accurately a mathematical model describes the true system in a real-life situation [16], is often a source of uncertainty. Moreover, model issues such as whether the training data and testing data are being generated by the same data distribution, or if the portion of the data universe that is provided to an algorithm in the training phase is substantially representative of the universe itself, bear a significant impact on the uncertainty involved in the system [17].
- **Algorithm Uncertainty:** Lastly, the algorithm of choice may often use numerical approximations that can result in uncertainty. Also, algorithm-related issues such as the suitability of the initial/boundary conditions of the system, or the choice of parameters in parametric methods, may add to this list of potential sources of uncertainty.

Approaches to Uncertainty Estimation

Over the years, several methods and theories have evolved to estimate/resolve uncertainty in pattern recognition. A broad categorization of these approaches is presented in Table 1.2.

Approach	Description
Probabilistic	The data is modeled as probability distributions, and the model outputs are computed as probabilities that capture the uncertainty. This is arguably the most popular approach, and used across various fields ranging from hydrology [18] to epidemiology [19].
Statistical	Uncertainty is estimated by analyzing the statistical properties of the model errors that occurred in reproducing observed data (as stated in [15]). The estimate is typically represented as a prediction interval (or a confidence interval), and is extensively used in statistics and machine learning.
Simulation/ Resampling-based	Methods such as Monte Carlo simulation use random samples of parameters or inputs to explore the behavior of a complex system or process, and thereby estimate the uncertainty involved [20]. This approach is once again widely used in financial modeling, robot localization, dynamic sensor networks and active vision [21].
Fuzzy	This approach, introduced by Zadeh [22], provides a non-probabilistic methodology to estimate uncertainty, where the membership function of the quantity is computed. This approach is widely used in consumer electronics, movie animation software, remote sensing and weather monitoring [23].
Evidence-based	Approaches such as the Dempster-Shafer theory [24], the more recent Dezert-Smarandache theory [25], possibility theory [26], and the MYCIN certainty factors [27] are approaches that are commonly used to resolve uncertainty when there are multiple evidences in the information fusion context.
Heuristic	Many approaches use application-specific heuristics or method-specific heuristics (such as measures based on the probability estimates produced by the k-Nearest Neighbor classifier [28], or ranking-based measures [29]) as the measure of uncertainty in the prediction.

Table 1.2: Categories of approaches to estimate uncertainty

Representations of Uncertainty Estimates

Just as there have been different approaches for estimating uncertainty, there have also been different representations of the estimate of a confidence measure (that captures the uncertainty). A categorization of these representations (largely inspired by the categorization presented by Langford[1]) is presented below:

- **Probability as Confidence:** This is easily the most common approach that is adopted universally by researchers that apply machine learning techniques to various applications. The probability of an event or occurrence is directly considered to be the confidence in the predicted result. It would be beyond the scope of this work to list all the earlier efforts that have adopted this approach, but a few examples can be found in [30], [31], [29], and [32]. Speech recognition is an example of an application domain where the posterior probability is popularly interpreted as the confidence.
- **Confidence Intervals:** Classical confidence intervals are most popular in statistics to convey an interval estimate of a parameter. Their usage in machine learning and pattern recognition has been relatively limited. Samples of earlier work where the uncertainty in pattern recognition models are represented as confidence intervals include the input space partitioning approach of Shrestha and Solomatine [15], the perturbation-resampling work of Jiang et al. with SVMs [33], Set Covering Machines by Marchand and Shawe Taylor [34], and the E^3 algorithm for learning the optimal policy in reinforcement learning by Kearns and Singh [35]. There are also variants of confidence intervals such as *asymptotic intervals* (approximate confidence intervals for

[1] http://hunch.net/?p=317

small samples, which become equivalent to confidence intervals when the number of samples increases) .

- **Credible Intervals:** Credible intervals [36] are also called *Bayesian confidence intervals*, since they are effectively the Bayesian 'subjective' equivalent of frequentist confidence intervals, where the problem-specific contextual prior information is incorporated in the computation of the intervals. Although this is treated as a separate category, the practical usage of credible intervals is often the same as confidence intervals. An example can be found in the work of Kuss et al. [37], where Markov Chain Monte Carlo (MCMC) methods are used to derive Bayesian confidence intervals of the posterior distribution in the analysis of psychometric functions.
- **Gamesman Intervals:** One of the earliest proponents of this approach is the new theory of conformal predictions proposed by Vovk, Shafer and Gammerman [38], where the prediction intervals/regions are based on game theory/betting contexts. The output prediction interval contains a set of predictions that contain the true output a large fraction of the time, and this fraction can be set by the user. (This approach is the basis of this dissertation work, and will be revisited in more detail later in the document).

Evaluating Uncertainty Estimates

A significant challenge for researchers in confidence estimation is the identification of appropriate metrics that can evaluate the obtained values. While there have been several approaches to overcoming this challenge, a few popular metrics are presented below:

- *Negative log probability:* Related efforts in the past [32] [39] have used the Negative Log Probability (NLP) as a metric of evaluating the 'goodness' of a

confidence measure. NLP is defined as:

$$NLP = \frac{-\sum_i \log p(c_i|x_i)}{n}$$

where c_is are the class labels in a classification problem. In regression, NLP is defined as:

$$NLP = \frac{-\sum_i \log p(y_i = t_i|x_i)}{n}$$

This metric is known to penalize both under-confident and over-confident predictions.

- *Normalized Cross Entropy:* Blatz et al. [32] pointed out that the NLP metric is sensitive to the base system's performance. To address this issue, they introduced the Normalized Cross Entropy (NCE) metric which measures the relative drop in log probability with respect to a baseline (NLP_b). NCE is given by:

$$NCE = \frac{NLL_b - NLL}{NLL_b}$$

- *Average Error:* This metric, representing the proportion of errors made over test data samples, is easily the most commonly used. This is defined as follows in the classification context [32]. Given a threshold τ and a decision function $g(x)$ which is equal to 1 when the classifier confidence measure is greater than τ, and 0 otherwise, the Average Classification Error (ACE) is given as:

$$ACE = \sum_i \frac{1 - \delta(g(x_i), c_i)}{n}$$

where δ is 1 if its arguments are equal, and 0 otherwise. In regression, this is defined as the Normalized Mean Square Error (NMSE):

$$ACE = \frac{1}{n} \sum_i \frac{(t_i - m_i)^2}{var(t)}$$

where t_is are the target predictions, and m_i is the mean of the predictive distribution $p(y_i|x_i)$.

- *ROC Curves:* Receiver Operating Characteristic (ROC) Curves [40] are also used in some cases to obtain a normalized view of the performance of classifiers and their confidence values.

In addition to the above metrics, there are several other metrics such as the LIFT loss [39] which have also been used in evaluating measures of confidence or uncertainty.

1.2 Understanding the Terms: Confidence and Probability

The terms 'confidence', 'probability', 'reliability', and 'belief' are often used interchangeably in the uncertainty estimation literature. There has been no explicit study or investigation to understand the usage of these terms, and it may not be possible to make conclusive statements about the meanings of any of these terms - since the choice of usage of these terms in earlier work has largely been application-driven or user-initiated, and hence, is largely subjective. However, a brief review of commonly accepted interpretations of the terms 'confidence' and 'probability', along with their commonalities and differences, is presented below.

Probability: The classical definition of the probability of an event (as defined by Laplace) is the ratio of the number of cases favorable to the occurrence of the event, to the number of all cases possible (when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible). However, there are several competing interpretations of the actual 'meaning' of probability values. Frequentists view probability simply as a measure of the frequency of outcomes (the more conventional interpretation), while Bayesians treat probability more subjectively as a statistical procedure that endeav-

ors to estimate parameters of an underlying distribution based on the observed distribution.

Mathematically, a probability measure (or distribution), P, for a random event, E, is a real-valued function, defined on the collection of events, F, defined on a measurable space Ω and satisfying the following axioms:

1. $0 \leq P(E) \leq 1 \forall E \in F$, where F is the event space, and E is any event in F.
2. $P(\Omega) = 1$ and $P(\emptyset) = 0$.
3. $P(E_1 \cup E_2 \cup \dots) = \sum_i P(E_i)$, if E_is are assumed to be disjoint.

These assumptions can be summarized as: Let (Ω, F, P) be a measure space with $P(\Omega) = 1$. Then (Ω, F, P) is a probability space, with sample space Ω, event space F and probability measure P. Note that the collection of events, F, is required to be a σ-algebra. (By definition, a σ-algebra over a set X is a nonempty collection of subsets of X, including X itself, which is closed under complementation and countable unions of its members).

Confidence: Formally, confidence can be written as a measurable function:

$$\Gamma : Z^* \times X \times (0,1) \rightarrow 2^Y$$

where Z is the set of all data-label pairs, X represents the new test data point, $(0,1)$ is the interval from which a confidence level is selected, and 2^Y is the set of all subsets of Y, the label space. However, while the label space in a classification problem is a finite set, the label space in regression problems is the real line itself.

If a user were to go by the mathematical definitions stated above, there is not much in common between confidence and probability, since the definitions clearly show them to be distinctly different. However, in common usage, these are often considered the same, and this has led to the thin line between the terms. With

both probability and confidence, there are frequentist and subjectivist (Bayesian) approaches. While the debate between these two schools of thought is more prominent with the usage of the term 'probability', confidence has two similar schools of thought too. These are represented as confidence intervals and Bayesian confidence intervals (or credible intervals). Classical confidence intervals are most popular in statistics to convey an interval estimate of a parameter. On the other hand, credible intervals are effectively the Bayesian 'subjective' equivalent of frequentist confidence intervals, where the problem-specific contextual prior information is incorporated in the computation of the intervals. The differences in the usages of these two terms can be viewed from two perspectives:

- The term 'confidence' is often associated with the concept of confidence intervals in statistics, which are interval estimates of a population parameter. In this context, 'confidence' of an estimate does not suggest the probability of the occurrence of the parameter estimate; rather, a range of estimates are together said to represent the confidence value. In fact, the confidence interval estimates indicate that if a value from the interval is chosen in the future, the number of errors can be restricted to $100-c\%$, where $c \in [0,100]$ is the confidence value. In common usage, a claim to 95% confidence in something is normally taken as indicating virtual certainty. In statistics, a claim to 95% confidence simply means that the researcher has seen something occur that only happens one time in twenty or less. This is very different from probability, as defined earlier in this section.
- From another technical perspective, probability is a measure associated with a particular random variable. Hence, the term probability is pertinent as long as the random variable is not observed. Once the observation is seen, there

is no more uncertainty, and the concept of probability is irrelevant. However, the confidence interval on the observation continues to provide an indication of the number of errors in future trials.

It may not be possible to make conclusions on which term is more relevant in a particular context, since there have been various perspectives to how these terms are used. As a cursory remark, it can be stated that probability values are most meaningful when the true distribution of the data is known. If not, it could be considered a more practical approach to provide confidence intervals and measures.

1.3 Confidence Estimation: Theories and Limitations

Although there have been several efforts to the computation of a confidence measure in pattern recognition (as mentioned earlier), each of them has its own advantages and limitations. In the following paragraphs, the limitations of existing approaches are presented, and a list of desiderata for a confidence measure is presented.

All approaches that provide confidence/probabilistic measures in machine learning algorithms that are used for pattern recognition (both classification and regression) and provide error guarantees can be broadly identified to be motivated by two theories, as stated in [41]. The two major theories are: *Bayesian Learning* and *Probably Approximately Correct (PAC) Learning*, each of which is discussed below.

Bayesian Learning

Without a doubt, Bayesian learning methods constitute the most popular approach to obtain probability values in pattern recognition applications. These methods are

based on the Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

where $P(A|B)$ is the posterior distribution, $P(B|A)$ is the likelihood, and $P(B)$ is the prior over the random variable B. A detailed review of Bayesian learning approaches can be found in [42], [43], and [10].

PAC Learning

PAC learning is a framework that was proposed by Valiant in 1984 [44] [45] to mathematically analyze the performance of machine learning algorithms. As stated in [46], "*in this framework, the learner receives samples and must select a generalization function (called the hypothesis) from a certain class of possible functions. The goal is that, with high probability (the probably part), the selected function will have low generalization error (the approximately correct part)*". In simpler words, the PAC learning approach is based on a formalism that can decide the amount of data required for a given classifier to achieve a given probability of correct predictions on a given fraction of future test data [47]. Given a collection of data instances X of length n, a set of target concepts C (class labels, for example), and a learner L using hypothesis space L:

> C is **PAC-learnable** by L using H if for all $c \in C$, distributions D over X, ε such that $0 < \varepsilon < \frac{1}{2}$, and δ such that $0 < \delta < \frac{1}{2}$, learner L will with probability at least $(1-\delta)$ output a hypothesis $h \in H$ such that $error_D(h) \leq \varepsilon$, in time that is polynomial in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$, n, and $size(C)$.

PAC theory has led to several practical algorithms, including boosting.

Limitations

Although the Bayesian and PAC learning approaches are used extensively in machine learning algorithms, the values generated by these algorithms are often impractical, invalid or unreliable. The limitations of these theories in obtaining practical reliable values of confidence are detailed in [41], [48], [49], [50] and [1], and are summarized below.

Bayesian learning approaches make a fundamental assumption on the probability distribution of the data. The values generated by Bayesian approaches are generally correct only when the observed data are actually generated by the assumed distribution, which does not happen often in real-world scenarios. When the data correctly corresponds to the assumed distribution, probability values generated by Bayesian algorithms are always valid. *Validity*, in this context, is defined as the correspondence of the probability value with the actual number of errors made with respect to the sample set, i.e. if the probability value is 0.73, there are exactly 27 errors if a similar data instance was picked from a data set of 100 instances. This property is also called *calibration*, and will be discussed later in this work.

Melluish et al. [1] conducted experiments to demonstrate this limitation of Bayesian methods when the underlying probability distribution of the data instances is not known. As shown in Figure 1.1, they showed that the number of errors made by the Bayesian ridge regression approach in the work varied as the a parameter was changed, which in turn modified the prior distribution. This directly illustrated the crucial role of the choice of the prior distribution to obtain valid measures of probability in Bayesian approaches.

In summary, the probability values obtained using Bayesian learning approaches face the following limitations:

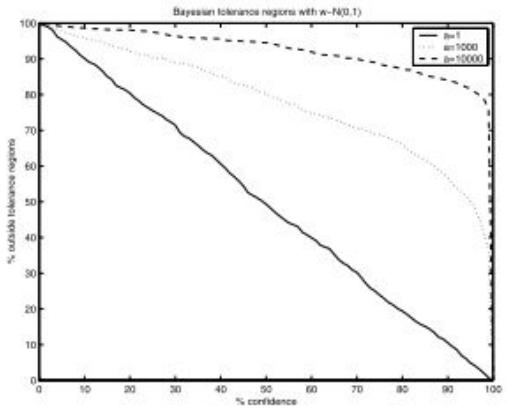


Figure 1.1: Bayesian tolerance regions on data generated with $\mathbf{w} \sim N(0,1)$. The figure plots the % of points outside the tolerance regions against the confidence level (Figure reproduced from [1])

- Such approaches have strong underlying assumptions on the nature of distribution of the data, and hence become invalid when the actual data in a problem do not follow the distribution.
- Many guarantees provided by the Bayesian theory are sometimes asymptotic, and may not apply to small sample sizes.

On the other hand, **PAC learning approaches** rely only on the i.i.d (identically independently distributed) assumption, and do not assume any other data distribution. However, the error bound values generated by such approaches are often not very practical, as demonstrated by Proedrou in [41], and by Nouretdinov in [51]. For example, Littlestone-Warmuth's Theorem is known to be one of the most sound results in PAC theory. The theorem states that for a two-class Support Vector Machine classifier f, the probability of mistakes is:

$$err(f) \leq \frac{1}{l-d}\left(d \ln \frac{el}{d} + \ln \frac{1}{\delta}\right) \tag{1.2}$$

with probability at least $1-\delta$, where $\delta \in (0,1]$, l is the training size, and d is the number of Support Vectors. For the USPS database from the UCI Machine Learn-

ing repository, the error bound given by this theorem for one out of ten classifiers (one for each of the digits) can be written as (the number of Support Vectors are 274 from [52]):

$$err(f) \leq \frac{1}{l-d}\left(d\ln\frac{el}{d}+\ln\frac{1}{\delta}\right) \approx \frac{1}{7291-274}274\ln\frac{7291e}{274} \approx 0.17 \qquad (1.3)$$

When extended to the ten classifiers, the error bound becomes 1.7, which is not practically useful. Nouretdinov also illustrated in [51] that the error bound becomes 0.74 when the Littlestone-Warmuth theorem is extended to multi-class classifiers for this dataset. In summary, the limitations of the PAC learning theory in the context of obtaining reliable confidence measure values are:

- The usefulness of the error bounds obtained is highly subjective, based on the dataset, classifier and the learning problem itself. There are settings where the error bounds are practically not useful.
- The obtained error bound values cannot be applied to individual test examples.

Given the limitations of existing theories, it becomes essential to identify and list the desired properties of confidence measures in machine learning applications.

1.4 Desiderata of Confidence Measures

A list of the desired features of 'ideal' confidence measures that are reliable and practically useful can be captured as follows:

1. *Validity:* Firstly, a confidence measure value should be *valid*, i.e. the number of errors made by the system is $1-t$, if the confidence value is given to be t. The measure is then said to be *well-calibrated*. In other words, the nominal coverage probability (confidence level) should hold, either exactly or to a good approximation [53].

2. *Accuracy:* The confidence measure value should bear a high positive correlation with the correctness of the prediction, i.e., an erroneous prediction should ideally have a low confidence value, and a correct prediction should typically have a high confidence value.

3. *Statistical Interpretation:* It would be useful if the confidence measure values obtained could be interpreted as confidence levels, as defined in traditional statistical models. This will allow seamless applications of mainstream statistical approaches in machine learning and pattern recognition, and vice versa.

4. *Optimality:* Given a confidence level, the methodology should construct prediction regions whose width is as narrow as possible.

5. *Generalizability:* The design of the computation methodology for the confidence measure should be generalizable to all kinds of classification/regression algorithms, and also applicable to multiple classifier/regressor systems.

1.5 Summary of Contributions

This dissertation contributes to the field of uncertainty estimation in multimedia computing by computing reliable confidence measures for machine learning algorithms that aid decision-making in real-world problems. Most existing approaches that compute a measure of confidence do not satisfy all the aforementioned desired features of such a measure. However, there have been recent developments towards a gamesman approach to the definition of confidence that satisfies many of the important properties listed above, including validity, statistical interpretation and generalizability. This theory is called the Conformal Predictions (CP) framework, and was recently developed by Vovk, Shafer and Gammerman [54] [38] based on the principles of algorithmic randomness, transductive inference and hypothesis

testing. This theory is based on the relationship derived between transductive inference and the Kolmogorov complexity [55] of an i.i.d. (identically independently distributed) sequence of data instances, and provides confidence measures that are well-calibrated. This theory is the basis of this work, and more details of the theory are presented in Chapter 2.1.

Confidence Estimation: Contributions

This dissertation applies the CP framework to multimedia pattern recognition problems in both classification and regression contexts. This work makes three specific contributions that aim to make the CP framework practically useful in real-world problems. These contributions, described in Chapters 3,5 and 6, are briefly summarized below.

1. Development of a methodology for learning a kernel function (or distance metric) that can be used to provide *optimal* and *accurate* conformal predictors.
2. Validation of the extensibility of the CP framework to multiple classifier systems in the information fusion context.
3. Extension of the CP framework to continuous online learning, where the measures of confidence computed by the framework are used for online active learning.

These contributions are validated using two classification-based applications (risk stratification in clinical decision support and multimodal biometrics), and two regression based applications (head pose estimation and saliency prediction in images). More details of these applications are presented in Chapter 2. In addition to the contributions mentioned above, other related contributions have also been

made as part of this dissertation in the respective application domains, and these are detailed in later chapters. A summary of these contributions is presented below.

1. Efficiency Maximization in Conformal Predictors: The CP framework has two important properties that define its utility, as defined by Vovk et al. [38]: *validity* and *efficiency*. As described in Chapter 2, validity refers to controlling the frequency of errors within a pre-specified error threshold, ε, at the confidence level $1-\varepsilon$. Also, since the framework outputs prediction sets at a particular confidence level, it is essential that the prediction sets are as small as possible. This property is called *efficiency*.

Evidently, an ideal implementation of the framework would ensure that the algorithm provides high efficiency along with validity. However, this is not a straightforward task, and depends on the learning algorithm (classification or regression, as the case may be) as well as the non-conformity measure chosen in a given context. In this work, a framework to learn a kernel (or distance metric) that will maximize the efficiency in a given context is proposed. More details of the approach and its validation are discussed in Chapters 3 and 4.

2. Conformal Predictions for Information Fusion: The CP framework ensures the calibration property in the estimation of confidence in pattern recognition. Most of the existing work in this context has been carried out using single classification systems and ensemble classifiers (such as boosting). However, there been a recent growth in the use of multimodal fusion algorithms and multiple classifier systems. A study of the relevance of the CP framework to such systems could have widespread impact. For example, when person recognition is performed with the face modality and the speech modality individually, how can these results be combined to provide a measure of confidence? Would it be possible to maintain the calibration property when there is multiple evidence, and these are fused at the

decision level? The details of this contribution are discussed further in Chapter 5.

3. Online Active Learning using Conformal Predictors: As increasing amounts of data are generated each day, labeling of data has become an equally increasing challenge. Active learning techniques have become popular to identify selected data instances that may be effective in training a classifier. All these techniques have been developed within the scope of two distinct settings: *pool-based* and *online (stream-based)*. In the pool-based setting, the active learning technique is used to select a limited number of examples from a pool of unlabeled data, and subsequently labeled by an expert to train a classifier. In the online setting, new examples are sequentially encountered, and for each of these new examples, the active learning technique has to decide if the example needs to be selected to re-train the classifier.

One of the key features of the CP framework is the calibration of the obtained confidence values in an online setting. Probabilities generated by traditional inductive inference approaches in an online setting are often not meaningful since the model needs to be continuously updated with every new example. However, the theory behind the CP framework guarantees that the confidence values obtained using this transductive inference framework manifest as the actual error frequencies in the online setting, i.e. they are well-calibrated [56]. Further, this framework can be used with any classifier or meta-classifier (such as Support Vector Machines, k-Nearest Neighbors, Adaboost, etc). In this work, we propose a novel active learning approach based on the p-values generated by this transductive inference framework. This contribution is discussed in more detail in Chapter 6.

Application Domains: Challenges and Contributions

The CP framework is most pertinent to risk-sensitive applications, where the cost of an error in the decision is high. It would be imperative in such applications to be able to control the frequency of errors committed. Medical diagnosis and security/surveillance applications are two such risk-sensitive applications, where an error may be very costly to the protection of human life (or lives). These application domains have been selected in this work to validate the three contributions in the classification setting. The other two applications are selected to validate the proposed contributions, when extended to the regression formulation.

A summary of the application domains used in this work is presented in Tables 1.3 and 1.4. More details of these application domains are presented in Chapter 2. In addition to the contributions based on the CP framework, there have been other contributions based on machine learning and pattern recognition that have been made, as part of this dissertation, towards solving the challenges in each of the applications. These contributions are also outlined in these tables.

1.6 Thesis Outline

The remainder of this dissertation is structured as follows. Chapter 2 is divided into two major sections: theory and application. Section 2.1 discusses the background of the Conformal Predictions framework, and its advantages and limitations. Section 2.2 presents the background of the application domains considered in this work, and also the corresponding datasets that have been used for all the experiments in this dissertation. Chapter 2 concludes with a study of the empirical performance of the Conformal Predictions framework. Chapters 3 and 4 present the proposed methodologies for maximizing efficiency in the CP framework for classification and regression respectively. Chapter 5 details our findings on applying the

Risk Prediction in Cardiac Decision Support (Classification)	
Problem description	∗ Classify a patient into one of two categories based on whether the patient is likely to face complications following a coronary stent procedure ∗ High risk-sensitivity ∗ Solution needs *validity* as well as high *efficiency*, to be useful
Proposed solution	An appropriate kernel function that can maximize efficiency within the CP framework, while maintaining validity, is learnt from the data
Other contributions	A clinically relevant inter-patient kernel metric has been developed combining evidence (using patient attributes) and knowledge (using the SNOMED medical ontology)

Head Pose Estimation for the Social Interaction Assistant (Regression)	
Problem description	∗ Estimate the head pose of an individual, independent of the identity, using face images ∗ In real-world scenarios, it may not be feasible to obtain the absolute pose angle using computer vision techniques. It would be a more practical approach to provide a region of possible head pose angle values, depending on a confidence level that the user chooses
Proposed solution	∗An appropriate distance metric that maximizes efficiency in the CP framework for regression, is learnt from the training data and labels ∗ A new framework for supervised manifold learning called Biased Manifold Embedding has been proposed, and this has been used for learning the required metric

Table 1.3: A summary of the applications and the corresponding contributions-I

Multimodal Person Recognition in the Social Interaction Assistant (Classification)	
Problem description	∗ Recognize an individual using both face and speech modalities, and associate reliable measures of confidence for multimodal person recognition results ∗ High risk-sensitivity in security/surveillance situations ∗ While there have been many existing efforts to estimate the confidence of recognition in each modality individually, the computation of confidence when there are two modalities involved is not as well-studied
Proposed solution	The decision obtained from each modality is considered as an independent statistical test, and the combination of p-values obtained from the CP framework is used to study the calibration of the final results
Other contributions	Online active learning algorithm using the CP framework has been proposed for face recognition. A batch mode active learning technique using numerical optimization, and a person-specific feature selection method have also been proposed to enhance performance in face recognition algorithms

Saliency Prediction in Images (Regression)	
Problem description	∗ Compute the saliency of regions in medical images (such as X-rays) during diagnosis, using eye gaze data of radiologists ∗ High risk-sensitivity ∗ Solution needs *validity* as well as high *efficiency*, to be useful ∗ Multiple image features may need to be used to determine saliency
Proposed solution	∗ A regression model is developed to predict saliency based on each relevant image feature. The result of each of these models is considered as an independent statistical test, and the combination of p-values obtained from the CP framework is used to study the calibration of the final results ∗ The CP framework is thus used to identify salient regions in the images, based on a specified confidence level
Other contributions	An integrated approach to combine top-down and bottom-up perspectives for prediction of saliency in videos has been proposed and implemented

Table 1.4: A summary of the applications and the corresponding contributions-II

CP framework to information fusion in both classification and regression settings, and Chapter 6 presents the novel Generalized Query by Transduction framework for online active learning that has been proposed based on the theory of Conformal Predictions. Chapter 7 summarizes the contributions and outcomes of this dissertation, providing pointers to directions of future work.

Chapter 2

BACKGROUND

This chapter lays down the background of this work from both theory and application perspectives. The chapter begins by describing the theory behind the Conformal Predictions framework, and the details of how it is used in both classification and regression contexts. From the application perspective, this chapter introduces the domains considered in this work, and describes the datasets used in this work.

2.1 Theory of Conformal Predictions

The theory of conformal predictions was recently developed by Vovk, Shafer and Gammerman [54] [38] based on the principles of algorithmic randomness, transductive inference and hypothesis testing. This theory is based on the relationship derived between transductive inference and the Kolmogorov complexity [55] of an i.i.d. (identically independently distributed) sequence of data instances. Hypothesis testing is subsequently used to construct conformal prediction regions, and obtain reliable measures of confidence.

If $l(Z)$ is the length of a binary string Z, and $C(Z)$ is its Kolmogorov complexity (the length of the minimal description of Z using a universal description language), then:

$$\delta(Z) = l(Z) - C(Z) \tag{2.1}$$

where $\delta(Z)$ is called the *randomness deficiency* of the string Z. This definition provides a connection between incompressibility and randomness. Intuitively, Equation 2.1 states that lower the value of $C(Z)$, higher the $\delta(Z)$, or the lack of randomness. The Martin-Lof test for randomness provides a method to connect randomness with statistical hypothesis testing. This test can be summarized as a function

$t : Z^* \rightarrow \mathbb{N}$ (the set of natural numbers with 0 and ∞), such that $\forall n \in \mathbb{N}, m \in \mathbb{N}, P \in P_n$:

$$P\{z \in Z^n : t(z) \geq m\} \leq 2^{-m} \tag{2.2}$$

where P_n is the set of all i.i.d. probability distributions. Equation 2.2 can also be written as:

$$P\{z \in Z^n : t(z) \in [m, \infty)\} \leq 2^{-m} \tag{2.3}$$

Now, if we use the transformation $f(x) = 2^{-x}$, Equation 2.3 can in turn be written in terms of a new function $t'(z)$:

$$P\left\{z \in Z^n : t'(z) \in (0, 1]\right\} \leq 2^{-m} \tag{2.4}$$

Hence, a function $t' : Z^* \rightarrow (0, 1]$ is a Martin-Lof test for randomness if $\forall m, n \in \mathbb{N}$, the following holds true:

$$P\left\{z \in Z^n : t'(z) \leq 2^{-m}\right\} \leq 2^{-m} \tag{2.5}$$

If 2^{-m} is substituted for a constant, say r, and r is restricted to the interval $[0, 1]$, Equation 2.5 is equivalent to the definition of a *p-value* typically used in statistics for hypothesis testing. Given a null hypothesis H_0 and a test statistic, p-value is simply defined as the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. In other words, the p-value is the smallest significance level of the test for which H_0 is rejected based on the observed data, i.e. the p-value provides a measure of the extent to which the observed data supports or disproves the null hypothesis.

In order to apply the above theory to pattern classification problems, Vovk et al. [38] defined a *non-conformity measure* that quantifies the conformity of a data point to a particular class label. This non-conformity measure can be appropriately designed for any classifier under consideration, thereby allowing the concept to

be generalized to different kinds of pattern classification problems. To illustrate this idea, the non-conformity measure of a data point x_i for a k-Nearest Neighbor classifier is defined as:

$$\alpha_i^y = \frac{\sum_{j=1}^{k} D_{ij}^y}{\sum_{j=1}^{k} D_{ij}^{-y}} \tag{2.6}$$

where D_i^y denotes the list of sorted distances between a particular data point x_i and other data points with the same class label, say y. D_i^{-y} denotes the list of sorted

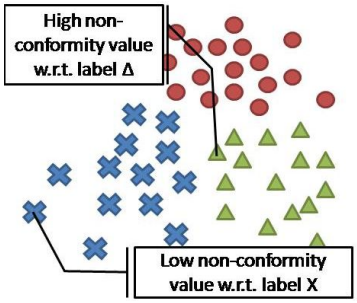


Figure 2.1: An illustration of the non-conformity measure defined for k-NN

distances between x_i and data points with any class label other than y. D_{ij}^y is the jth shortest distance in the list of sorted distances, D_i^y. In short, α_i^y measures the distance of the k nearest neighbors belonging to the class label y, against the k nearest neighbors from data points with other class labels (Figure 2.1). Note that the higher the value of α_i^y, the more non-conformal the data point is with respect to the current class label i.e. the probability of it belonging to other classes is high.

The methodologies for applying the Conformal Predictions (CP) in classification and regression settings are described in the following subsections.

Conformal Predictors in Classification

Given a new test data point, say x_{n+1}, a null hypothesis is assumed that x_{n+1} belongs to the class label, say, y_p. The non-conformity measures of all the data points in the system so far are re-computed assuming the null hypothesis is true. A p-value

function (which satisfies the Martin-Lof test definition in Equation 2.5) is defined as:

$$p(\alpha_{n+1}^{y_p}) = \frac{count\left\{i : \alpha_i^{y_p} \geq \alpha_{n+1}^{y_p}\right\}}{n+1} \tag{2.7}$$

where $\alpha_{n+1}^{y_p}$ is the non-conformity measure of x_{n+1}, assuming it is assigned the class label y_p. In simple terms, Equation 2.7 states that the p-value of a data instance belonging to a particular label is the normalized count of the data instances that have a higher non-conformity score than the current data instance, x_{n+1}. It is evident that the p-value is highest when all non-conformity measures of training data belonging to class y_p are higher than that of the new test point, x_{n+1}, which points out that x_{n+1} is *most conformal* to the class y_p. This process is repeated with the null hypothesis supporting each of the class labels, and the highest of the p-values is used to decide the actual class label assigned to x_{n+1}, thus providing a transductive inferential procedure for classification. If p_j and p_k are the two highest p-values obtained (in respective order), then p_j is called the *credibility* of the decision, and $1-p_k$ is the *confidence* of the classifier in the decision. The p-values

Algorithm 1 Conformal Predictors for Classification

Require: Training set $T = \{(x_1,y_1),...,(x_n,y_n)\}$, $x_i \in X$, number of classes M, $y_i \in Y = y_1,y_2,\ldots,y_M$, classifier Ξ

1: Get new unlabeled example x_{n+1}.
2: **for** all class labels, y_j, where $j = 1,\ldots,M$ **do**
3: Assign label y_j to x_{n+1}.
4: Update the classifier Ξ, with $T \cup \left\{x_{n+1},y_j\right\}$.
5: Compute non-conformity measure value, $\alpha_i^{y_j} \forall i = 1,\ldots,n+1$ to compute the p-value, P_j, w.r.t. class y_j (Equation 2.7) using the conformal predictions framework.
6: **end for**
7: Output the conformal prediction regions $\Gamma_{1-\varepsilon} = \left\{y_j : P_j > \varepsilon, y_j \in Y\right\}$, where $1-\varepsilon$ is the confidence level.

generated using this approach satisfy the modified Martin-Lof test in Equation 2.5.

The conformal prediction regions are presented as regions representing a specified confidence level, Γ_ε, which contain all the class labels with a p-value greater than $1-\varepsilon$. These regions are *conformal* i.e. the confidence threshold, $1-\varepsilon$ directly translates to the frequency of errors, ε in the online setting [54, 56]. The approach is summarized in Algorithm 1. The CP framework can be used in association with any classifier, with the suitable definition of a non-conformity measure. Sample non-conformity measures for various classification algorithms are presented below in Table 2.1.

Conformal Predictors in Regression

The CP framework has also been used in regression formulations to deliver prediction regions that are calibrated [38] [48] [41] [66]. While the label space in a classification problem is a finite set, the label space in regression problems is the real line itself. This needs a different methodology of applying the framework, since it is not pragmatic to hypothesize each value on the real line as a possible class label, and compute a corresponding p-value. The algorithm to define conformal prediction regions for regression seeks to identify intervals (or neighborhoods) on the real line that conform to a pre-specified confidence level. Evidently, a larger confidence level (say $1-\varepsilon_1$) will result in a larger interval $\Gamma_{1-\varepsilon_1}$ that ensures the required confidence, and a smaller confidence level (say $1-\varepsilon_2$) will result in a narrower interval $\Gamma_{1-\varepsilon_2}$. It should be noted here that $\Gamma_{1-\varepsilon_2} \subseteq \Gamma_{1-\varepsilon_1}$, as long as $1-\varepsilon_2 \leq 1-\varepsilon_1$.

In a regression problem, the non-conformity measure of a data-label entity, say (x,y), can be defined as the absolute value of the difference between y and the predicted value, $\hat{y}$, calculated from x and the old (training) examples [38] (Equation 2.8).

$$\alpha_i = |y_i - \hat{y}_i| \tag{2.8}$$

Classifier	Non-conformity measure	Description
k-NN	$\frac{\sum_{j=1}^{k} D_{ij}^{y}}{\sum_{j=1}^{k} D_{ij}^{-y}}$	Ratio of the sum of the distances to the k nearest neighbors belonging to the same class as the hypothesis y, and the sum of the distances to the k nearest neighbors belonging to all other classes [38] [57] [58].
Support Vector Machines	Lagrange multipliers, or $e^{-ad_i^m}$	A suitable function of the distance of a data point from the hyperplane [38] [59] [60] [61].
Neural networks	$\frac{\sum_{y' \in Y : y' \neq y} o_{y'}}{o_y + \gamma}$	Ratio of the sum of the output values of all output neurons except the winning neuron and the output value of the winning neuron itself. γ is a parameter that can be varied [38] [62] [49] [63] [64].
Logistic regression	$\begin{cases} 1+\exp^{-w.x} & \text{,y=1} \\ 1+\exp^{w.x} & \text{,y=0} \end{cases}$	Reciprocal of the estimated probability of the observed y given the observed x for a given data instance. w is the weight vector typically computed using Maximum Likelihood Estimation [38].
Boosting	$\sum_{t=1}^{T} \alpha_t B_t(x,y)$	Weighted sum of the individual non-conformity measures of each of the weak classifiers B_t, and α_t are the weights learnt by the boosting algorithm [38]
Random forests	$\frac{out_{raw} - \overline{out_{raw}}}{\sigma}$, where $out_{raw}(i) = \frac{nsample}{\overline{p(i)}}$, and $\overline{p(i)} = \sum_j \lvert prox(i,j) \rvert^2$	Scaled outlier measure of an observed x with respect to label $y \in Y$, and other data instances belonging to the same class [65]. $nsample$ is the number of samples in the class under consideration, and $prox(i,j)$ is the similarity between two data instances in a random forest.

Table 2.1: Non-conformity measures for various classifiers

Papadopoulos et al. [48] also suggested a modified non-conformity measure where the predicted accuracy of the decision rule f on a training set is used, i.e. the measure is defined as:

$$\alpha_i = \frac{|y_i - \hat{y}_i|}{\sigma_i} \tag{2.9}$$

where σ_i is an estimate of the accuracy of the decision rule f on x_i.

An efficient algorithm to compute conformal prediction intervals in the case of ridge regression (regularized least squares regression) was proposed by Nouretdinov et al. [66], and is described below in Algorithm 4. For more details of the method, please refer Chapter 4 of this dissertation or [66].

Algorithm 2 Conformal Predictors for Regression

Require: Training set $T = \{(x_1, y_1), ..., (x_n, y_n)\}$, new example x_{n+1}, confidence level r, $X = x_1, x_2, \ldots, x_{n+1}$

1: Calculate $C = I - X(X'X + \alpha I)^{-1}X'$ (for ridge regression).
2: Let $A = C(y_1, y_2, \ldots, y_n, 0)' = (a_1, a_2, \ldots, a_{n+1})$
3: Let $B = C(0, 0, \ldots, 0, 1)' = (b_1, b_2, \ldots, b_{n+1})$
4: **for** $i = 1$ to $n+1$, do **do**
5: Calculate u_i and v_i.
If $b_i \neq b_{n+1}$, then $u_i = min(\frac{a_i - a_{n+1}}{b_{n+1} - b_i}, \frac{-(a_i + a_{n+1})}{b_{n+1} + b_i})$; $v_i = max(\frac{a_i - a_{n+1}}{b_{n+1} - b_i}, \frac{-(a_i + a_{n+1})}{b_{n+1} + b_i})$
If $b_i = b_{n+1}$, then $u_i = v_i = \frac{-(a_i + a_{n+1})}{2b_i}$.
6: **end for**
7: **for** $i = 1$ to $n+1$, do **do**
8: Compute S_i according to Equation 2.10 below.
9: **end for**
10: Sort $(-\infty, u_1, u_2, \ldots, u_{n+1}, v_1, \ldots, v_{n+1}, \infty)$ in ascending order, obtaining $\hat{y_0}, \ldots, \hat{y_{2n+3}}$
11: Output $\cup_i [\hat{y_i}, \hat{y_{i+1}}]$, such that $N(\hat{y_i} > r$, where $N(y_i) = \#S_j : [\hat{y_i}, \hat{y_{i+1}}] \subseteq S_j$, where $i = 0, \ldots, 2n$, and $j = 1, \ldots, n+1$.

S_i in Algorithm 4 is given by the following equation:

$$S_i = \begin{cases} [u_i, v_i] & \text{if } b_{n+1} > b_i \\ (-\infty, u_i] \cup [v_i, \infty) & \text{if } b_{n+1} < b_i \\ [u_i, \infty) & \text{if } b_{n+1} = b_i > 0 \text{ and } a_{n+1} < a_i \\ (-\infty, v_i] & \text{if } b_{n+1} = b_i > 0 \text{ and } a_{n+1} > a_i \\ \Re & \text{if } b_{n+1} = b_i = 0 \text{ and } |a_{n+1}| \leq |a_i| \\ \Phi & \text{if } b_{n+1} = b_i = 0 \text{ and } |a_{n+1}| > |a_i| \end{cases} \tag{2.10}$$

In addition to being used in classification and regression formulations, the CP framework has also been used for feature selection, where an optimization problem is formulated to minimize the strangeness (non-conformity) values [67]. This approach was found to be effective in bio-informatics for microarray classification. In an interesting recent work, Hardoon et al. [68] proposed a model selection strategy using non-conformity measures (as defined in the CP framework) that is comparable to traditional strategies such as Cross-Validation and Leave-One-Out, but has theoretical guarantees for success and faster convergence rates. They demonstrated their work using SVMs.

Assumptions and Their Impact

The main (and only) assumption of the CP framework is that data should be i.i.d. (identically independently distributed). This is defined as the *randomness assumption* in the framework. There have been several efforts in recent years that have focused on developing machine learning frameworks that are based on only this assumption on the data. One example is the statistical learning theory of Vapnik and Chervonenkis [69], whose manifestation is the Support Vector Machine algorithm. All the theoretical results of the CP framework (for example, the frequency of errors, ε, is always under a specified confidence threshold 1- ε) hold under this i.i.d. (or randomness) assumption.

However, many of the results in the framework can actually hold with an even weaker assumption on the data model called exchangeability. Exchangeable distributions have also been used in machine learning, and the most popular example is the bag-of-words modeling assumption in natural language processing. An exchangeable distribution P means that for every positive integer n, every permutation π of $1,2,\ldots,n$:

$$P(z_1,z_2,\ldots) \in Z^\infty : (z_1,z_2,z_n) \in E = P(z_1,z_2,\ldots) \in Z^\infty : \\ (\pi(z_1),\pi(z_2),\ldots,\pi(z_n)) \in E$$

where $z_i \in Z : X \rightarrow Y$, where X is the data samples and Y are the labels. In simpler terms, an exchangeable sequence of samples is a sequence that future samples behave similar to samples that have already been observed, i.e. any order of samples is equally likely in the observed sequence[1]. Every i.i.d. sequence is exchangeable, but not vice versa (for example, sampling without replacement is exchangeable, but not i.i.d.). Exchangeability implies that variables have the same distribution. On the other hand, exchangeable variables need not be independent [70]. Most of the results in the CP framework hold under this exchangeability assumption.

Impact of Assumptions: To study the impact of the above assumption on the CP framework, a brief study to test the exchangeability of a data stream was carried out using the USPS data set from the UCI Machine Learning repository [71]. Vovk et al. [38] proposed a methodology to test the exchangeability of the observed data sequence in an online manner. This methodology was based on the definition of an exchangeability supermartingale. The usual statistical approach to testing (sometimes called the Neyman-Pearson-Wald theory) is essentially offline. However, this approach is online, i.e., we constantly update the strength of evidence against the

[1] http://en.wikipedia.org/wiki/Exchangeable_random_variables

null hypothesis of randomness. A brief introduction to martingales is presented below.

Martingales: A martingale is a stochastic process such that the conditional expected value of an observation at some time t, given all the observations up to some earlier time s, is equal to the observation at that earlier time s. A discrete-time martingale can be defined as a stochastic process $X_1, X_2, X_3, \ldots$ such that for all n, the following condition is satisfied:

$$E(X_{n+1}|X_1, \ldots, X_n) = X_n$$

For the continuous-time equivalent, the martingale needs to satisfy the following condition:

$$E(Y_t|X_T, T \leq s) = Y_s, \forall s \leq t$$

A (discrete-time) submartingale is a sequence $X_1, X_2, X_3, \ldots$ satisfying the condition:

$$E(X_{n+1}|X_1, \ldots, X_n) \geq X_n$$

and a supermartingale satisfies the condition:

$$E(X_{n+1}|X_1, \ldots, X_n) \leq X_n$$

Testing for Exchangeability: The methodology of using martingales to test for exchangeability was proposed by Vovk in [38]. After observing a new data point, a learner outputs a positive martingale value reflecting the strength of evidence found against the null hypothesis of data exchangeability. Finding evidence against the null hypothesis is equivalent to gambling against it, and the strength of evidence equals the gambler's current capital. More details of this approach to testing are presented in [72]. Vovk inferred that there existed a family of 'exchangeability

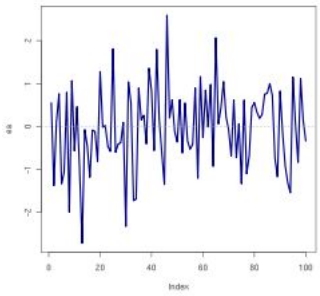


Figure 2.2: Example of a martingale sequence

martingales', which can be successfully applied to detecting lack of randomness. These are called the randomized power martingales, and are given by:

$$M_n^{\varepsilon} = \prod_{i=1}^{n} \varepsilon p_i^{\varepsilon-1}$$

where p_is are the p-values provided by the CP framework, and $\varepsilon \in [0,1]$ is a parameter that can be varied. As the martingale is evaluated for each new data example in an online fashion, a higher value (tending towards infinity) suggests that the exchangeability condition is violated, and a lower value (tending to zero) suggests that the data stream is indeed exchangeable.

Experiments with the USPS dataset: The exchangeability test was applied to the UCI USPS dataset, on which the CP framework has been shown to demonstrate valid results in earlier work [56] (when the data instances are randomly permuted). The CP framework implemented in this study was based on the *k*-Nearest Neighbor (*k*-NN), to replicate the settings of [56]. When the randomized power martingale (RPM) is applied to the USPS dataset (as is, i.e. the data is not permuted), the result is shown in Figure 2.3. This figure shows that the value of the RPM keeps increasing as more examples are added. This suggests that the dataset is not exchangeable, since the evidence collected against the null hypothesis of exchangeability is extremely high.

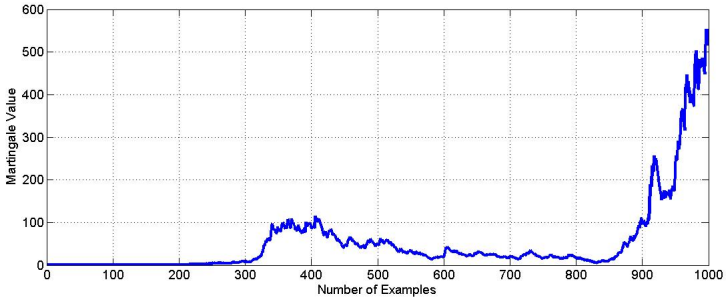


Figure 2.3: Randomized power martingale applied to the USPS dataset. It is evident that this dataset is not exchangeable

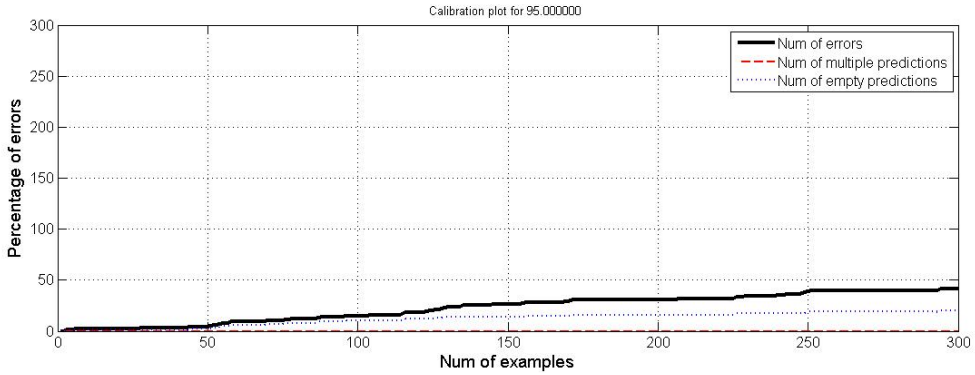


Figure 2.4: Results of the CP framework using kNN at the 95% confidence level. Note that the number of errors are far greater than 5%, i.e., the CP framework is not valid in this case

The results of applying the CP framework using k-NN (as considered above) on the non-permuted USPS dataset at the 95% confidence level is presented below. Evidently, the frequency of errors is far more than 5%, thus establishing that the framework does not provide valid results, since the dataset is not exchangeable (or i.i.d, which is a stronger assumption). On the other hand, when the USPS dataset is randomly permuted, the results of the randomized power martingale are shown below in Figure 2.5. This figure shows that the randomly permuted USPS dataset is exchangeable. Note that the RPM tends to zero, as more examples are added. As mentioned earlier, it was observed in [56] that the CP framework provided valid

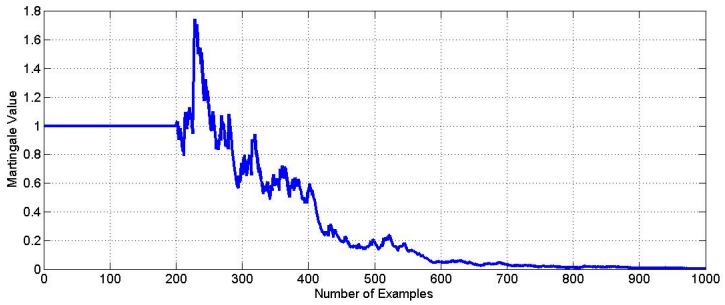


Figure 2.5: Randomized power martingale applied to the randomly permuted USPS dataset. Notice that the data is now exchangeable, since the RPM tends to zero, as more examples are added

results for such a randomly permuted USPS dataset. The above study clearly illustrates the impact of the exchangeability assumption on the CP framework. When this assumption is not satisfied on the data stream, the validity property is affected, i.e., there can be no guarantee provided on the frequency of errors made by the framework at a given confidence level. However, since random permutation addresses this issue, batch learning problems can conveniently be permuted to yield valid measures of confidence.

Advantages, Limitations and Variants

The desirable properties of the CP framework are summarized by Vovk et al. [38] (in Chapter 1 of their book). We briefly review these properties, before discussing the limitations, and the variants of the framework that have been introduced to offset these limitations.

- *Validity:* This can be stated to be the most important property of the framework. The conformal predictors are always *valid*, i,e. the frequency of errors does not exceed a pre-specified error threshold, ε, at every confidence level $1-\varepsilon$. This is often also called the *calibration* property.

- *Efficiency:* Since the framework outputs prediction sets at a particular confidence level, it is essential that the prediction sets are as small as possible. This property is called *efficiency*. In case of classification, this would be equivalent to having the least possible number of class labels in the prediction set. In regression, the predicted interval must be as narrow as possible.

- *Nested Prediction Sets:* The output prediction set at a lower confidence level is a subset of the prediction set at a higher confidence level, leading to nested prediction sets. For example, as stated earlier, $\Gamma_{0.65} \subseteq \Gamma_{0.9}$. Why is this important? On one hand, this helps statistically interpret the predictions, since this property is similar to those of confidence intervals (as stated in Chapter 1). On the other hand, this provides for an intuitive presentation of results in machine learning, at large.

- *Conditionality:* The output prediction set is constructed with complete consideration of the current example being observed. This is a very essential property for computing measures of confidence, since such frameworks (especially frameworks such as PAC learning which provide error bounds) often do not consider the current observed data instance in the computations.

- *Generalizability:* (called *flexibility* in [38]) The framework is extensible to any kind of machine learning algorithms for classification and regression, as long as a suitable non-conformity measure is defined. Thus, if a particular algorithm is suitable for an application (say neural networks), this framework can be applied on top of the algorithm to obtain conformal prediction regions as the output.

The framework has some limitations too, which are discussed below. These have been elaborated in [38] (Chapter 4), and variants of the framework that address

these limitations have also been proposed. Two of these limitations are mentioned below.

- *Computational inefficiency:* A major limiting factor of the framework (in its transductive form) is the seeming computational inefficiency of the framework. Since the framework is based on transductive inference, the non-conformity measure has to be recomputed for all the data instances when a new data instance enters the system. This is a huge computational overhead. This resulted in the design of the Inductive Conformal Predictors framework [38] [48] [63], where the training set is divided into training and calibration portions. The calibration portion is used to compute the p-values when a new data example is observed, thus significantly reducing the required computations. However, this approach trades off computational efficiency for a loss in predictive efficiency (as defined earlier in this sub-section), and hence has to be implemented after careful empirical evaluation.
- *Conditional Validity:* Conformal predictors are not automatically conditionally valid, i.e., data belonging to a particular class may be more difficult to recognize than other data entities. Hence, it is natural to expect that at the 95% confidence level, the error rate will be significantly greater than 5% for the difficult classes; validity only ensures that the average error rate over *all* class labels will be close to 5%. The notion of Mondrian conformal predictor is introduced to address this concern. For more details of this approach, please refer [38].

2.2 Application Domains and Datasets Used

The contributions in this dissertation have been validated on problems from four different application domains, representative of variety in the challenges to be ad-

dressed. These application domains were chosen to validate the contributions in both classification and regression contexts. These problems are listed below, and each of them is further described in the following subsections.

- Risk prediction in cardiac decision support (Classification)
- Head pose estimation for a Social Interaction Assistant to help individuals with visual impairments (Regression)
- Multimodal person recognition for the Social Interaction Assistant (Classification)
- Saliency prediction in radiological images (Regression)

Risk Prediction in Cardiac Decision Support

Machine learning algorithms such as Support Vector Machines [73], genetic algorithms [74], and neural networks [75] [76] have been used in cardiology to improve the quality of care, stratify risk, and provide prognostications. Traditional learning algorithms learn from data of past patients, and provide predictions on new patients, without convincing information of the reliability or confidence in the predictions. In medical diagnosis/prognosis, it is extremely essential to evaluate the performance of such algorithms on the risk of possible error in supporting the decision-making process. In this work, the CP framework has been used to achieve this purpose. Unlike many conventional classification systems, this framework allows us not just to risk classify new patients, but add *valid* measures of confidence in our predictions for every *individual* patient. The objective of the work in this application domain is to predict the risk of complications following a coronary Drug Eluting Stent procedure (DES), using patient data provided by Advanced Cardiac Specialists, a cardiology practice based in Arizona, USA.

(a) Balloon Angioplasty (b) Bare Metal Stents (c) Drug Eluting Stents

Figure 2.6: Percutaneous Coronary Intervention procedures for management of Coronary Artery Disease (CAD)

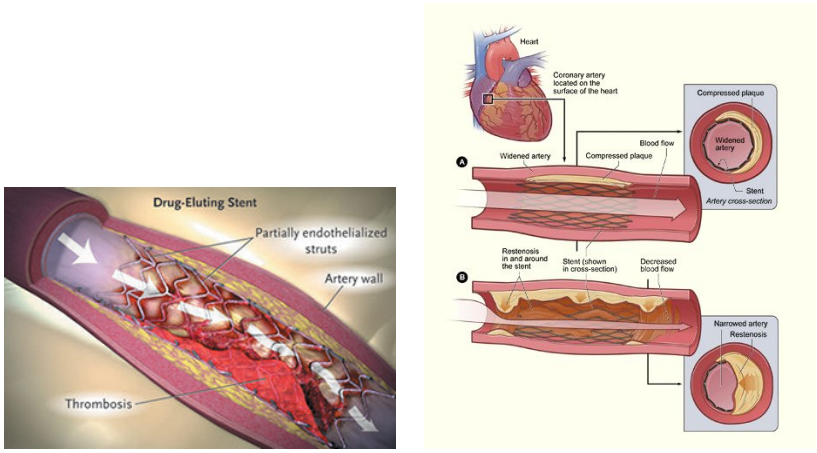
Drug Eluting Stents (DES) have emerged as the de facto option for Percutaneous Coronary Intervention (PCI), with distinct advantages over bare metal stents [77]. Since restenosis rates are less than 10% with DES, there has been an explosive growth in their use over a very short period. However, unanticipated complications have been increasingly observed following a coronary DES procedure. In addition to standard Major Adverse Cardiac Events (MACE) and procedural complications associated with all PCI procedures, DES have resulted in additional complications, including late Stent Thrombosis, increased incidence of early Stent Thrombosis, and late restenosis, which could result in myocardial infarction (heart attack) or death.

The predictive model proposed in this work helps to stratify the risk for a specific patient for post-DES complications, and thereby stratify patient populations according to healthcare requirements, reducing the need for unnecessary invasive procedures with their attendant risks and significant costs. The *valid* measures of confidence can be used by the physician to make an informed, evidence-based decision to manage a patient, choosing the most appropriate option from repeat PCI, Coronary Artery Bypass Graft surgery (CABG), and/or maximized medical therapy to minimize the possibility of occurrence/recurrence.

CATEGORY	ATTRIBUTES
Demographic and Clinical Presentation	Age, Gender, Ejection Fraction, Diabetes, Hypertension, Hyperlipidemia, Smoking, Race, Acute Coronary Syndrome (Acute MI, Unstable Angina), Chronic Stable Angina, Cardiogenic Shock, Congestive Heart Failure, Pulmonary Edema
History	Previous Myocardial Infarction (Acute MI, Silent MI), Unstable Angina, Chronic Stable Angina, Previous PCI, Previous CABG, Previous Stroke, Cardiogenic Shock, Congestive Heart Failure
Angiographic	Vessel, No. of Lesions treated, Bifurcation lesion, Narrowed Coronary Arteries, Multi-vessel Disease, Target Coronary Artery (Left Anterior Descending, Diagonal Left Circumflex, Obtuse Marginal, Right Posterolateral,Right Posterior Descending, Saphenous vein Graft), Coronary Lesion Characteristics (Calcific, Eccentric, Diffuse Disease, Ostial Disease, Total Occlusion, Thrombus), Vessel Tortuosity, Reference Vessel Diameter, Lesion Length, Restenotic lesion, Lesion Type (A, B1, B2, C), Thrombus, Pre-procedure TIMI = 0
Procedural	Urgent/Emergent, Balloon Predilatation (Diameter, Length, Balloon to artery ratio, Maximal Predilatation Inflation Pressure), Stent Implantation (Stent Length, Diameter, 2.25 mm stent, Stent length / Lesion length ratio, Maximal Stent balloon inflation pressure), Postprocedure TIMI flow < 3, Left main Stenting, Multiple stents, Dissection, Acute reocclusion

Table 2.2: Patient attributes used in the Cardiac Patient dataset

Data Setup: Data was obtained from the central Percutaneous Coronary Intervention registry maintained at Advanced Cardiac Specialists (ACS), consisting of patient cases across the state of Arizona (including cases of different genders, races and ethnic groups). 2312 patient cases who had a DES procedure performed during the period 2003 to 2007, and who had followed up with the cardiac care facility during the 12 months following the procedure, were selected from the PCI registry as the dataset for the development of the model. The complications considered for this model included: Stent Thrombosis and Restenosis, which manifest as chest pain,



(a) Stent Thrombosis (b) Stent Restenosis

Figure 2.7: Complications following a Drug Eluting Stent (DES) procedure

myocardial infarction and sometimes even death. All patient particulars including demographics, clinical parameters, patient history, angiographic, procedural and follow-up details (a total of 165 patient attributes) were obtained as available in the registry. These attributes are listed in Table 2.2. The dataset was extracted as a Comma Separated Value (CSV) format file from the PCI registry which was maintained in SPSS. All patient data was handled in compliance with the U.S. Food and Drug Administration's (FDA) Protection of Human Subjects Regulations 45 CFR (part 46) and 21 CFR (parts 50 and 56) and the U.S. Department of Health and Human Services Health Insurance Portability and Accountability Act (HIPAA) of 1996.

The data was cleaned and missing values were handled in the most clinically relevant manner, where appropriate. The data was subsequently normalized. Of the selected patient cases, only 182 (only 7.87% of the total data) had a complication at 12 months following DES. To handle class imbalance (approximately, 92% to 8%) in the patient data, our experiments illustrated the effectiveness of the Synthetic Minority Over-sampling Technique (SMOTE) [78] to obtain good performance with

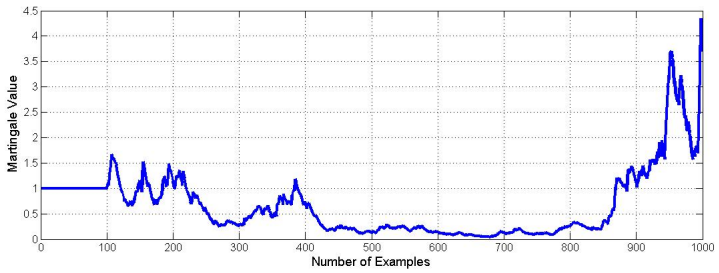


Figure 2.8: Results of the randomized power martingale on the non-permuted cardiac patient data stream. Note that this figure is inconclusive; the martingale value does not tend towards infinity, nor towards zero

imbalanced data. All steps of data extraction, pre-processing, and model development were carried out in MATLAB R2007b. The SVM-KM toolbox [79] was used for the algorithm implementation. The results of this work are discussed further in Chapter 3.

Testing Exchangeability: As mentioned in the previous section, the only assumption for the CP framework to provide valid results is that the data should be i.i.d.; rather, the data should be exchangeable (a weaker assumption, as stated earlier) i.e. the order in which the data samples arrive is random and can be permuted. To study the validity of the assumption for this dataset, a randomized power martingale (described in Chapter 1) is constructed and used to test the exchangeability of the data. We present the results of the randomized power martingale (without random permutation of the data) in Figure 2.8. Note that the martingale value is very low (around 4-5), and hence, makes the study inconclusive. Since the martingale does not tend towards infinity or towards zero, it is not possible to state conclusively about the exchangeability of the dataset.

However, when the dataset is randomly permuted, the results obtained are presented in Figure 2.9. Note that the martingale tends towards zero in this case. This figure shows that when the data stream is randomly permuted, the exchangeability

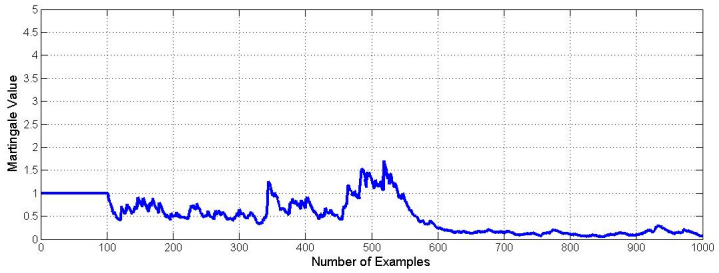


Figure 2.9: Results of the randomized power martingale on the randomly permuted Cardiac Patient dataset. Note that the martingale tends towards zero

condition is satisfied. It is acceptable to assume that the data can be randomly permuted before applying the CP framework in this application (even in the real-world setting), since the patient data that has already been collected and stored can conveniently be permuted before using the CP framework. Hence, the guarantees of the CP framework will hold true in this application.

Head Pose Estimation in the Social Interaction Assistant

Head pose estimation has been studied as an integral part of biometrics and surveillance systems for many years, with its applications to 3D face modeling, gaze direction detection, and pose-invariant person identification from face images. With the growing need for robust applications, face-based biometric systems require the ability to handle significant head pose variations. In addition to being a component of face recognition systems, it is important to determine the head pose angle from a face image, independent of the identity of the individual. This can be of significant use in applications ranging from driver monitoring to 3D face recognition. While coarse pose angle estimation from face images has been reasonably successful in recent years [80], accurate person-independent head pose estimation from face images is a more difficult problem, and continues to elicit effective solutions.

The Social Interaction Assistant: The objectives of this work are anchored on

Figure 2.10: A first wearable prototype of the Social Interaction Assistant

the design and development of an assistive system that can help individuals with visual impairments in daily social interactions, called the *Social Interaction Assistant*. People who are blind are often at a disadvantage in daily interactions, for they are not aware of the presence of people around them, their identities or where they are looking. A group of researchers at the Center for Cognitive Ubiquitous Computing (CUbiC) at Arizona State University have been working towards a solution for this problem [81] [82] [83]. This dissertation forms a significant component of these efforts.

In order to identify unmet needs of the visually impaired community, two focus groups consisting primarily of people who are blind, as well as disability specialists and parents of students with visual impairment and blindness, were engaged in studies to understand their needs. During these focus groups, the participants agreed on many issues as being important problems. However, one particular problem - that of engaging freely with their sighted counterparts - was highlighted as a particularly important problem that was not being addressed by technology specialists. This led to the conceptualization of a wearable assistive device (Figure 2.10) that would allow a person who is blind or visually impaired to interact with sighted peers without those peers even being aware of their disability, or their assistive device, and this device was called the Social Interaction Assistant [81]. The focus

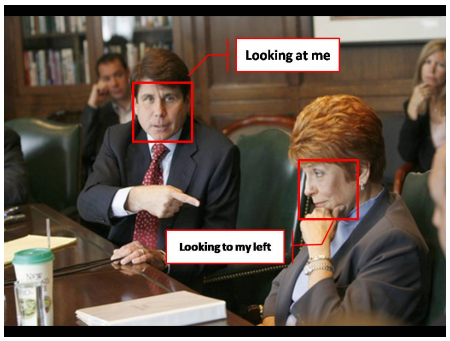


Figure 2.11: A sample application scenario for the head pose estimation system

group studies were used to identify and enumerate a list of needs for people who are blind, as they engage in social interactions, and one of these important needs was identified to be '*knowing where a person is directing his/her attention*'. Head pose estimation is an important element in providing a useful interaction experience for individuals who are blind. Further, the person recognition module can be triggered when it is known that an individual is looking at the user, i.e. the head pose estimation module can be used along with the person recognition module for a more practically useful device.

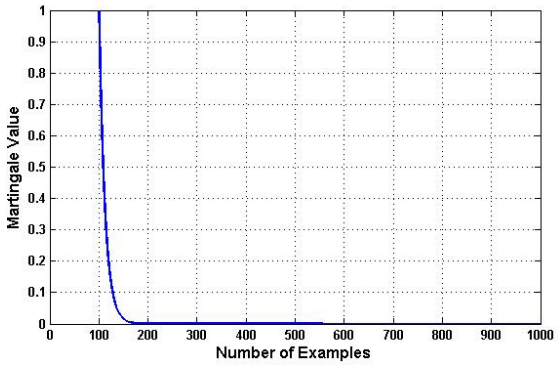
Data Setup: The FacePix database [84], built at the Center for Cognitive Ubiquitous Computing (CUbiC), has been used in this work for experiments and evaluation. Earlier work on head pose analysis have used databases such as FERET, XM2VTS, the CMU PIE Database, AT & T, Oulu Physics Database, Yale Face Database, Yale B Database and MIT Database for evaluating the performance of algorithms. Some of these databases provide face images with a wide variety of pose angles and illumination angles. However, none of them use a precisely calibrated mechanism for acquiring pose and illumination angles. To achieve a precise measure of recognition robustness, FacePix was compiled to contain face images with pose and illumination angles annotated in 1 degree increments.

Figure 2.12: Sample face images with varying pose and illumination from the FacePix database

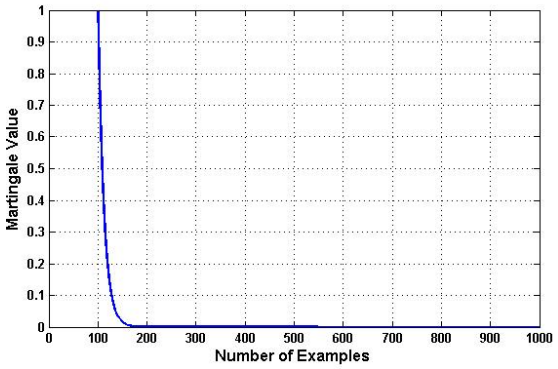
The FacePix database consists of three sets of face images: one set with pose angle variations, and two sets with illumination angle variations. Each of these sets are composed of a set of 181 face images (representing angles from $-90\,^{\circ}$ to $+90\,^{\circ}$ at 1 degree increments) of 30 different subjects, with a total of 5430 images. All the face images (elements) are 128 pixels wide and 128 pixels high. These images are normalized, such that the eyes are centered on the 57th row of pixels from the top, and the mouth is centered on the 87th row of pixels. The pose angle images appear to rotate such that the eyes, nose, and mouth features remain centered in each image. Also, although the images are down sampled, they are scaled as much horizontally as vertically, thus maintaining their original aspect ratios. Figure 2.12 provides two examples extracted from the database, showing pose angles and illumination angles ranging from $-90\,^{\circ}$ to $+90\,^{\circ}$ in steps of $10\,^{\circ}$. For earlier work using images from this database, please refer [84]. This database is publicly available[2], and has been used earlier by other researchers for head pose estimation [85] [86].

Testing Exchangeability: Similar to the previous subsection, a randomized power martingale was constructed and used to test the exchangeability of this dataset. The results of the martingale for randomly permuted face images from the FacePix

[2]http://www.facepix.org/



(a) Grayscale pixel intensity feature space



(b) Laplacian of Gaussian feature space

Figure 2.13: Results of the randomized power martingale when applied to the randomly permuted FacePix data

dataset are presented in Figure 2.13. Both the grayscale pixel intensity and the Laplacian of Gaussian feature spaces (which were used in this work) were studied. Note that the martingale values tend towards zero, establishing that both the feature spaces, when randomly permuted, are exchangeable, and hence, well-suited for our work with the CP framework.

Multimodal Person Recognition in the Social Interaction Assistant

In the wave of growing concerns about security and privacy, the need to reliably estimate the identity of an individual has become very pronounced. This has motivated active research in the field of biometrics. Biometric systems rely on the evi-

dence provided by face, voice, fingerprint, signature and other modalities to verify and validate the identity claimed by an individual. Modalities such as fingerprints and iris have proven to be very robust when the cooperation of the human subject can be assumed, both during enrollment and during test. This makes them ideal for limiting entry into secured areas (such as buildings) to known and trusted individuals. However, these biometrics are not very useful for recognizing people in public places, where there is little or no motivation to cooperate with the system.

The development of an assistive person recognition system for people who are blind provides a more tractable problem for face recognition researchers than security and surveillance applications [87]. It imposes a somewhat less stringent set of requirements because:

- the number of people to be recognized is generally smaller,
- disguise is not a serious concern,
- multiple pose angles, facial expressions and speech tones of a person can be captured as training images (unlike datasets in security or surveillance, where face images of miscreants typically contain only frontal and profile views of each persons face, with no intermediate views), and
- the person recognition process is a collaboration between the system and the user.

Moreover, focus group studies [81] that were conducted indicated that an important feature that blind users expected in a Social Interaction Assistant is the ability to *know the identities of the people standing in front of them*. In this dissertation, a multimodal approach using the *face* and *speech* modalities is adopted towards the development of an assistive person recognition system. Both of these modalities

are non-intrusive and 'at-a-distance', i.e. they can be used without the user necessarily having to pass through a sensing device (unlike fingerprint/iris-based person recognition).

Why a Multimodal Approach? Most biometric systems used in real world applications are unimodal [88]; that is, they rely on a single modality to carry out the authentication task. Such systems suffer from a variety of problems:

- the data collected may be corrupted by noise,
- a user may interact incorrectly with a sensor, for example can provide an incorrect facial pose,
- it is possible that a particular trait of two different persons are very similar, or
- a single trait may be subject to spoof attacks.

Multimodal systems seek to alleviate some of these problems by consolidating the evidence from multiple sensors. This can lead to better and reliable performance of the recognition/validation system. The individual pieces of information being fairly independent, are more robust to noise. In the context of this work, while each modality (video and audio) is limited by certain environmental conditions, such as changes in ambient lighting or background noise, by combining the two modalities we increase the viability of the system and the range of environments in which it can operate. Multimodal biometric systems can be classified into five categories as shown in Figure 2.14 [2]. In this classification, the present work can be categorized as a *Multiple Biometric System* approach, which combines the use of multiple modalities.

In face and speech-based biometrics, there are many sources of uncertainty, such as variations in pose, illumination and expressions in face images, or varia-

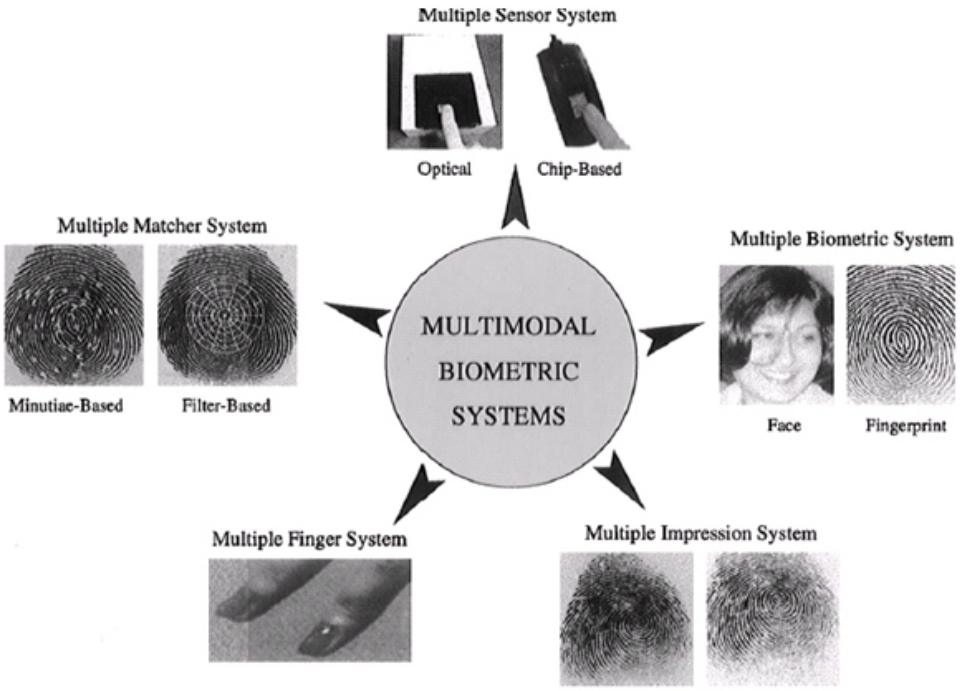


Figure 2.14: Categorization of approaches towards multimodal biometrics (Illustration reproduced from [2])

tions in tonality and pitch in speech. Considering that algorithms in the field of biometrics have direct implications in security and surveillance, it is essential that such algorithms provide a reliable measure of confidence on the predicted identity (or identities). Further, in the current application context, users who are blind have expressed their need to have the ability to interpret the reliability of the results obtained from a person recognition device. Obtaining valid confidence measures in multimodal person recognition constitutes the objective of this work.

Data Setup: The VidTIMIT [89] and the MOBIO (Mobile Biometry)[3] datasets are used to validate the proposed contributions. Both these databases contain frontal images of subjects under natural conditions, and simulate the scenario of a visually impaired individual in daily interactions. The VidTIMIT dataset contains the video

[3]http://www.mobioproject.org

recordings of 42 subjects reciting short sentences. The MOBIO (Mobile Biometry) dataset was created for the MOBIO challenge to test the performances of state-of-the-art face and speech recognition algorithms. It contains videos of 160 subjects captured using a mobile phone camera under challenging real world conditions. More details of the data capture can be found in [90].

For both these datasets, automated face cropping was performed to crop out the face regions [91] (In the VidTIMIT dataset, each of the videos were first sliced and stored as JPEG images of resolution 512 by 384). To extract the facial features, block based discrete cosine transform (DCT) was used (similar to [92]). Each image was subdivided into 8 by 8 non-overlapping blocks, and the DCT co-efficients of each block were then ordered according to the zigzag scan pattern. The DC co-efficient was discarded for illumination normalization, and the first 10 AC co-efficients of each block were selected to form compact local feature vectors. Each local feature vector was normalized to unit norm. Concatenating the features from the individual blocks yielded the global feature vector for the entire image. The cropped face image had a resolution of 128 by 128 and thus the dimensionality of the extracted feature vector was 2560. Principal Component Analysis (PCA), a commonly accepted step in face recognition techniques, was then applied to reduce the dimension to 100, retaining about 99% of the variance.

The speech data components of the VidTIMIT and MOBIO datasets were processed and handled by our collaborators at Tecnologico de Monterey, Mexico. More details of the corresponding speech data processing techniques used can be found in [93].

Testing Exchangeability: Similar to the previous application, a randomized power martingale was constructed and used to test the exchangeability of this dataset. The results of the martingale for face images of 5 subjects from the VidTIMIT

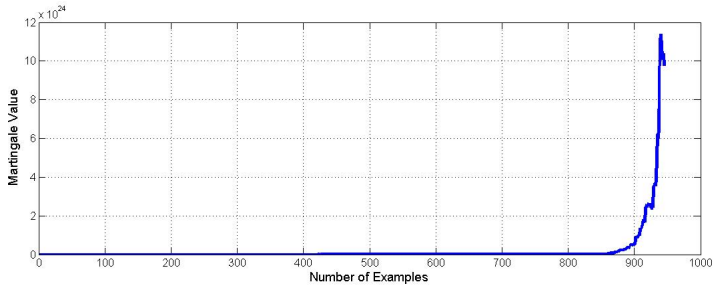


Figure 2.15: Results of the randomized power martingale with the VidTIMIT dataset. The data was not permuted. Note that it is clearly evident that the dataset is not exchangeable.

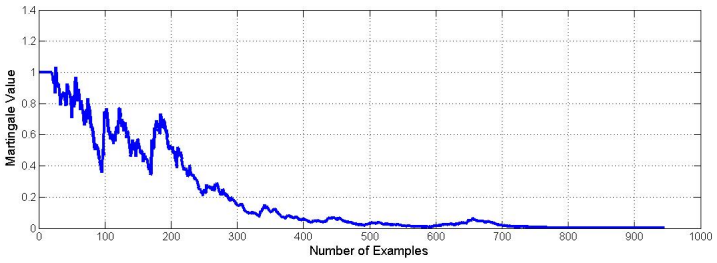
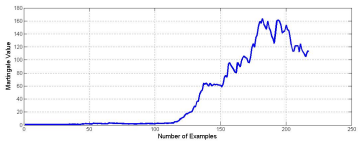
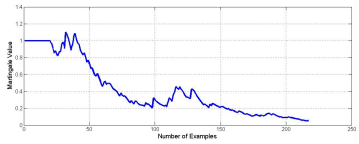


Figure 2.16: Results of the randomized power martingale with the randomly permuted VidTIMIT dataset. Note that it is clearly evident that the martingale tends towards zero, establishing that the permuted data is exchangeable

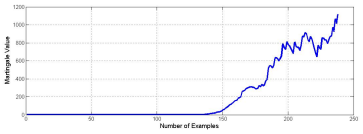
dataset are presented in Figure 2.15. Note that the martingale value tends towards infinity, establishing that the dataset, in its non-permuted form, is not exchangeable. However, when the same data is randomly permuted, we obtain the results shown in Figure 2.16. It is evident that the randomly permuted data is exchangeable, and hence, well-suited for our work with the CP framework. However, it is also possible in this application that a particular test setting may contain the data of just one subject (unlike the previous experiment, where the data from all the subjects was used to construct the martingale). For example, the video of a particular subject may have been recorded during a session, and provided as input to the recognition framework. To understand the exchangeability of the data generated from only one



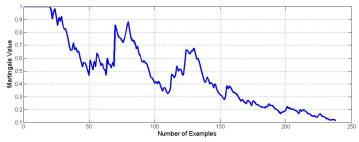
(a) Non-permuted data stream of Subject 1



(b) Randomly permuted data stream of Subject 1



(c) Non-permuted data stream of Subject 2



(d) Randomly permuted data stream of Subject 2

Figure 2.17: Results of the randomized power martingale when applied to the data stream of a single user

subject, we performed the same study with the face images of a single subject. We repeated this study for 2 different subjects, and the results of this study are presented in Figure 2.17. Evidently, the data stream, as generated, is not exchangeable; but once the stream is randomly permuted, the data stream is exchangeable.

Saliency Prediction in Radiological Images

Saliency is defined as a measure of possible user interest on a single unit (or pixel) of an image. In this work, we adopt a machine learning approach to learn saliency in a given application using the regions of interest indicated by human eye gaze using eye-tracking technology. The work is validated on radiological images viewed by radiologists, where it is valuable to learn from eye gaze information of expert radiologists and suggest possible regions of high saliency to help novice radiologists (or for that matter, any other user). When the user specifies a chosen confidence level, the system identifies regions of saliency that are conformal with respect to this confidence level.

As growing numbers of images and videos are generated each day, there has

been an equally increasing need to reliably identify appropriate regions of interest for all analysis tasks, such as medical diagnosis and surveillance. Radiological images constitute a special class of images that are used for a very specific purpose (i.e. diagnosis) and their *'correct'* interpretation is vitally important to patients. New radiologists are trained over a period of years to interpret radiological images, with a learning process that involves daily interactions with experienced radiologists. However, even after years of training, errors are not uncommon. Renfrew et al. [94] noted that such errors typically involved:

- lesions that were outside the area of interest in an image,
- a lack of knowledge,
- a failure to continue searching for abnormalities after the first abnormality was found, and
- failure to recognize a normal biologic variant.

This raises the question of whether sophisticated image analysis and machine learning techniques could be used to assist radiologists, by directing their attention to regions of images that might be of particular importance. Such tools might be especially helpful in high fatigue and stress scenarios, and in satisfaction of search error scenarios, which are known to be a major source of error in medical image analysis [95] [96] [97].

Eye-tracking is the procedure of tracking the position of eye gaze of a user. One of the earliest uses of eye-tracking was in the field of psychology in understanding how text is read. Researchers analyzed the variations in fixation and saccade durations with line spacing and difficulty of textual content. Eye-tracking was also used to understand scene and art perception. In more recent times, eye-tracking is

being increasingly used in various commercial and research applications ranging from Human Computer Interaction (HCI) and medical research to marketing.

Eye tracking technology has been used to study the nature of expertise in radiology, and to compare experts to novices [98] [99]. In 2002 [100], Dempere-Marco et al. analyzed the eye tracking data of two experienced radiologists while reading CT images of lungs. Their analysis was based on the spatial location of fixations, and the time spent at each fixation. The fixations were spatially clustered, and textural features were extracted from each cluster. Factor analysis was then used to find features that might be useful in a decision support system. In 2003, Hu et al. [101] and later in 2007, Antonelli et al. [102] also analyzed the visual fixations of experienced radiologists, with the motivation of providing decision support for medical imaging. As the radiologists studied a set of medical images, their scan paths were mapped into a feature domain, where the distribution of fixations was very different from that in the spatial domain. Specifically, the scan path was projected into a textural feature space that was spanned by the same textural parameters used in [100]. The 'hot spots' were then identified in the textural space. Each hot spot represented a particular combination of textural parameters that tended to attract and hold the attention of the radiologists. The next step was to back project those hot spots from the feature space into the spatial domain, to identify the spatial regions of interest. The last step was to select the 'most consistent' spatial regions of interest across all of the images. The result of this last step was a map showing the spatial regions of the images that were presumably more important to radiologists. Using this data, we propose a methodology to learn visual saliency in radiological images using human eye movements. More details of the methodology and our experiments are presented in Chapter 5.

Data Setup: A desktop Tobii[4] 1750 eye tracker with an LCD monitor (1280

Figure 2.18: Tobii 1750 eye tracker

Rad.	Specialty	Chest X-rays/week	Title	Years
01	Cardiothoracic	500	Staff	15
02	MRI	15	Fellow	1
03	Cardiothoracic	250	Staff	3
04	Musculoskeletal	15	Staff	6
05	Thoracic	200	Staff	10

Table 2.3: Participants' demographical information

x 1024 resolution) was used to record eye movement. This device, as shown in Figure 2.18, is integrated into a 21 inch monitor. It tracks the eye gaze of the viewers while using the monitor with a sampling frequency of 50 Hz. This eye tracker has a nominal accuracy of 0.5 degrees with moderate head movement.

Five radiologists (4 males and 1 female) from Mayo Clinic in Scottsdale, Arizona participated in an experiment. (For simplicity, the male pronoun 'he' will be used to represent all participants in the following discussion.) All 5 had normal or corrected-to-normal vision. Table 2.3 summarizes the demographic data collected from these 5 radiologists. This demographic data was used to estimate a heuristic expertise level for each radiologist as follows:

$$E = RPW + S(T + Y) \tag{2.11}$$

where, E is the expertise level, RPW is the average number of chest x-rays read per week, S is the specialty (0.9 if a cardiothoracic specialist, and 0.1 otherwise),

[4]http://www.tobii.com

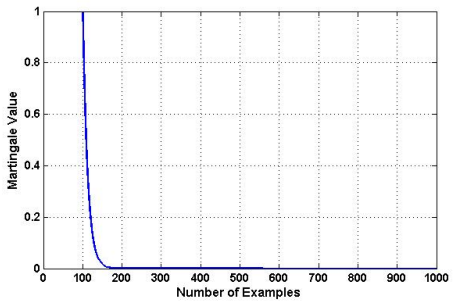
T is the title (1 for a staff, and 0 otherwise), and Y is the number of years of radiological experience. This expertise representation was formed in consultation with the radiologists at the Mayo Clinic, and was intended to be used to weight the inputs of the different radiologists appropriately.

Eye tracking data were recorded as these radiologists each read 20 chest x-ray images for diagnostic purposes. Some of these x-ray images were normal, and others were abnormal. No clinical history was provided with these x-ray images. For more details about the experimental procedure, please refer [98]. This dataset is used for learning a saliency predictor in Chapter 5.

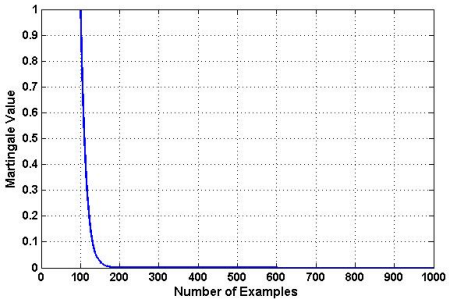
Testing Exchangeability: As was done with other datasets, a randomized power martingale was constructed and used to test the exchangeability of this dataset. In a preliminary study [98], four different feature spaces - Localized Edge Orientation Histograms, Haar Wavelets, Gabor Filters, and Steerable Filters - were found to provide the best results towards effective saliency prediction. The results of the martingale for randomly permuted data from the Radiology dataset for each of these four feature spaces are presented in Figure 2.19. Note that the martingale values tend towards zero for all the feature spaces, thus establishing that this dataset, when randomly permuted, is exchangeable and hence suitable for our work.

2.3 Empirical Performance of the Conformal Predictions Framework: A Study

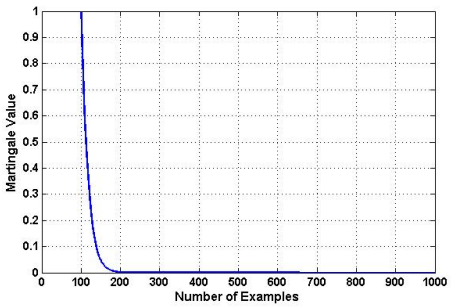
The validity of conformal predictors is known, and results demonstrating validity have been presented by several researchers over the last few years, as reviewed earlier. Results of validity in the application domains considered in this work have also been presented in Chapters 3, 4 and 5. In a different kind of study presented here, different uncertainty estimation frameworks have been analyzed against specific issues in learning systems that can be considered to affect system performance



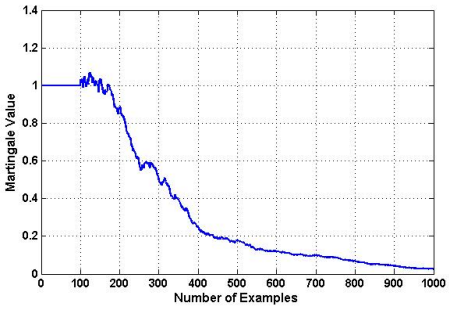
(a) Local Edge Orientation Histogram feature space



(b) Gabor Filters feature space



(c) Haar Wavelets feature space



(d) Steerable Filters feature space

Figure 2.19: Results of the randomized power martingale when applied to the randomly permuted Radiology dataset for each of the 4 feature spaces that were found to provide the best performances for effective saliency prediction

- error margins, training bias, data typicalness, and sample sizes. In particular, this study focuses on the relationship between confidence and correctness of a prediction in a classification setting. This study includes three different frameworks in its scope: Transductive Confidence Machines (TCM) [38] [41] (synonymous with the CP framework), Transductive Reliability Estimation (TRE) [103] [104] and a probabilistic approach based on boosting (called BP in this section, for convenience) [105] [106]. While the TCM approach has been detailed earlier in this chapter, the TRE and probabilistic approaches are briefly discussed below.

Transductive Reliability Estimation: Given two sets representing the data, S and $S \cup W$ (where W is the new data point), the ideal prediction for a test data point,

x_{test}, could be assumed to made when the data point is included in the training set. Hence, if we represented a trained model for a dataset as M_S, then the difference between predictions on x_{test} using models M_S and $M_{S\cup W}$ could be considered as providing a measure of the reliability of the prediction. This is the main idea of this framework.

To achieve this is a three-step process:

1. Obtain a probabilistic distribution on the output labels with a classifier, M_S, *induced* by the training data.
2. Include the new test data point, x_{test}, into the training set, and obtain a *transduced* model for $M_{S\cup W}$, and use this classifier to obtain a probabilistic distribution on the output labels. This transductive model is obtained using the CP framework (described earlier).
3. Compare the two probabilistic distributions using the normalized symmetric Kullback-Leibler divergence to obtain the reliability measure:

$$J_N(P,Q) = 1 - 2^{-\sum_{i=1}^{n}(p_i - q_i)\log\frac{p_i}{q_i}}$$

For more details on this approach, please refer Kukar's work [103] [104].

Probabilistic Approach: Since probabilistic approaches have different formulations, we have chosen one approach which aligns along a common implementation platform for all of these 3 frameworks. Adaboost.M2 [105] is used as a meta-learner, with k-Nearest Neighbors (k-NN) in each of its iterations as the base algorithm in this study. The probability distribution of k-NN on the output labels is equivalent to a Naive Bayes classifier using non-parametric density estimation with variable window sizes [106]. Also, the weighted addition of these probability values over the iterations in the Adaboost.M2 framework has been interpreted as

a Bayesian integration of probability values. This weighted sum of the probability values over the iterations is used as the measure of confidence, as often used in related work.

Experimental Setup

Dataset, Feature Spaces and Learning Algorithms: Related studies performed in earlier work often use synthetically generated data. Instead, we used a well-defined formulation of the head pose estimation problem using face images from the widely accepted FERET database [107] as the basis of our empirical study. We have defined the task in the problem as training a learning system to automatically classify captured face images into one of 3 labels - *frontal*, *left*, and *right*. The FERET database is widely respected as the standard database to evaluate face analysis algorithms. In this work, 4500 images with varying pose angles of different people were randomly selected from the FERET database. The ground truth for the class labels of these images was obtained using the FERET file nomenclature, where the last two letters of the file name of each image indicates the pose angle of the face in the image. The images were selected from the FERET database in such a manner that each of the 3 class labels - *frontal*, *left* and *right* had 1500 images. Since many of these face images include non-face information too, these images were manually cropped before our study. To obtain face images that are reasonably close to such a real-world setting for our analysis, a real-time face detection algorithm based on patch classifiers [108] was applied on images from the FERET face database. Samples of face portions extracted from the FERET database which were used for our analysis, are shown in Figure 2.20.

From these face portions, three different features were extracted:

- The pixel intensity values were used as is.

(a) (b) (c) (d)

Figure 2.20: Examples of face images from the FERET database (a and c) and the corresponding extracted face portions (b and d) used in our analysis

- Edges were extracted out of the image by thresholding the magnitude of the gradient computed from the vertical and horizontal Sobel filters, δ_y and δ_x respectively. The orientation of each of these edge pixels is computed as:

  $$\theta = \tanh(\frac{\delta_y}{\delta_x})$$

  Subsequently, a histogram of the orientations of these pixels is constructed with bins spanning the interval $[-180^\circ, +180^\circ]$. Initial experiments were carried out with 6, 8 and 12 bins with a regular k-NN classifier to study the performance, and a histogram with 12 bins was found to be most suitable.

- Gabor wavelet features at three different scales ($\{1,2,4\}$) and three different orientations ($\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$) were extracted from the image, and concatenated.

The Adaboost.M2 algorithm was used as a meta-learner [105]. In each of the iterations of Adaboost.M2, a k-NN (with a value of $k = 10$, chosen empirically) classifier was implemented on each of the 3 feature spaces independently, and the feature with the least error is selected as the weak learner of the iteration. The same learner was used in the TRE and TCM algorithms. For the TCM algorithm, the non-conformity measure was computed as specified in Table 2.1.

Experiment Design and Methodology: The objective of this study is to identify statistical factors in learning frameworks that can affect the confidence of a

Factor/Issue	Related Question
Data typicalness	How typical of the training points of the predicted class label is the given test data point?
Training bias	Was there a bias in the size/representativeness of the training dataset of the predicted class label?
Error margins	How close is the result value of the given algorithm (used for predicting using possible thresholds) to an ideal/mean result value obtained for data points with the predicted class label in the training phase?
Sample sizes	What portion of the possible universe of datasets was provided for training?

Table 2.4: A listing of factors pertinent to the evaluation of confidence estimation frameworks

learning system, and understand how existing confidence estimation frameworks address these factors. Table 2.4 lists a selected set of factors of evaluation, and the related questions that explain the corresponding factors. The listing is based on an intuitive understanding of factors that often cause issues in system performance. While this is not an exhaustive listing, nor can the factors be proved to be mutually independent, it was decided to perform this study with these factors since no earlier work with this motivation was performed before, and a study would be required to identify the factors themselves. It should also be mentioned that factors based on the performance in the training phase (often used in earlier related work) are represented by the data typicalness factor in this study.

Based on the factors listed in Table 2.4, the experiments designed and studied in this work are listed in Table 2.5.

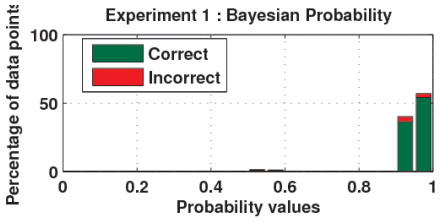
Results and Discussion

The results of each of the experiments in Table 2.5 are presented and discussed in this section. In this study, the accuracies obtained in the experiments are not given much attention. Instead, the focus is directed towards the *probability* value

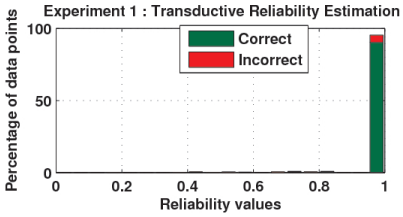
Index	Experiment Design	Notes/Factor(s) Addressed
1	1000 images each of *frontal*, *left* and *right* face images are used for training, and 500 images each of the 3 class labels are used for testing	Baseline study
2	1000 images each of *frontal*, *left* and *right* face images are used for training, but *left* and *right* face images with a pose angle of 22.5° are not included in training; Only *left* and *right* face images with pose angle 22.5° are used for testing	Study data typicalness with data that could be ambiguous
3	1000 images each of *frontal* and *left* are used for training, along with only 100 images of *right* face images, and the same 500 images of each class label as in Experiment 1 are provided for testing. Subsequently, the number of *right* face images for training is incremented in steps of 100 to study changes in performance	Study training bias
4	100 images each of *frontal*, *left* and *right* face images are used for training, and the same 500 images of each class label as in Experiment 1 are used for testing. Subsequently, the number of training images is incremented in steps of 100 to study changes in performance	Study effect of sample size relative to available universe of data

Table 2.5: Design of experiments for confidence measures in head pose estimation

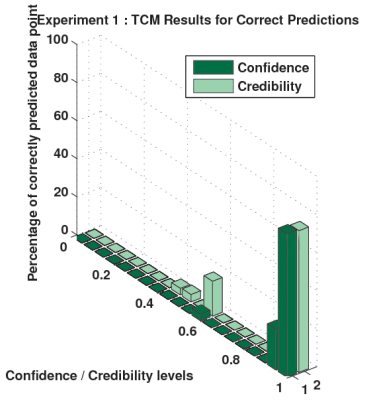
obtained from the BP method, the *confidence* and the *credibility* values obtained from the TCM framework approach, and the *reliability* value obtained from the TRE framework. An essential part of the analysis includes the study of the confidence values against correct and incorrect predictions. It is assumed that a better confidence estimation framework would result in low confidence values for incorrect predictions. The rationale behind this analysis is that if these frameworks were successful at declaring a low confidence on incorrect results, a threshold of confidence value could be used to generate better effective performance of a learning system. Looking at the results from Experiment 1 (Figure 2.21), it is interesting



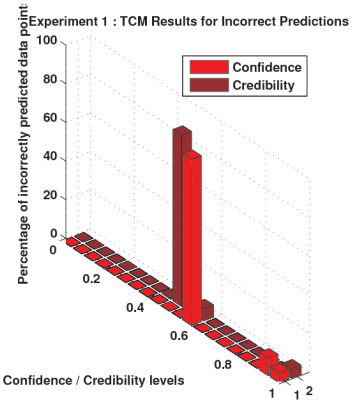
(a) Bayesian Probability



(b) Transductive Reliability Estimation



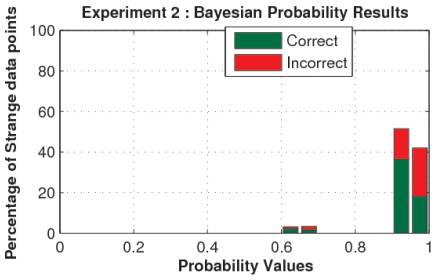
(c) TCM for correct predictions



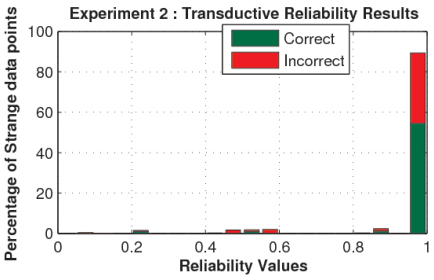
(d) TCM for incorrect predictions

Figure 2.21: Results of Experiment 1

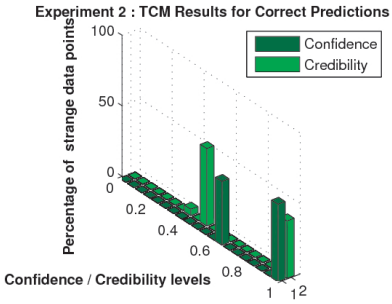
to note that the BP approach values tend to crowd around 0.9 to 1.0 for both correct and incorrect predictions. While the high accuracy of the approach makes the number of incorrect predictions look small, the fact that the system gives a high probability value on even incorrect predictions does not seem encouraging. A similar observation can be made for the TRE results, and in fact, in this case, almost all predictions have an extremely high reliability between 0.95 and 1.00. On the contrary, results from the TCM approach clearly indicate a difference between confidence/credibility values for correct and incorrect predictions. Incorrect predictions have confidence/credibility values in the 0.5-0.6 range, which suggests that this can possibly be thresholded to filter incorrect results. However, by the very definition of *confidence* and *credibility*, a low confidence implies a high credibility value, and a low credibility value implies a high confidence. Considering these issues, it would



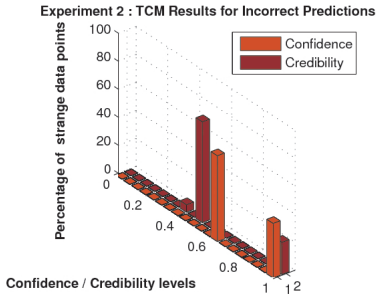
(a) Bayesian Probability Formulation



(b) Transductive Reliability Estimation



(c) TCM for correct predictions



(d) TCM for incorrect predictions

Figure 2.22: Results of Experiment 2

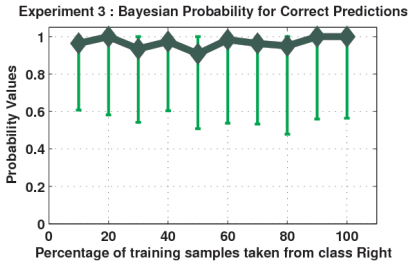
be necessary to understand the correlation between these two values, and deduce rules that can provide a final interpretation from the end user perspective.

The results from Experiment 2 (Figure 2.22) follow a trend that is very similar to Experiment 1. Here again, as Figure 2.22 indicates, while BP and TRE results still associate high confidence value with incorrect predictions, TCM provides objective assessments with significantly lower confidence/credibility even on correct predictions, which gives the end user the idea that the presented data stream may be untypical of the data used to train the system. An interesting observation in these results is that when 90° pose angle images are used for testing, the system manages to classify with ease (although these images were not provided during training); but the system is not as confident when 22.5° pose angle images are used, as shown by the percentage of incorrect predictions. While both these kinds of images are

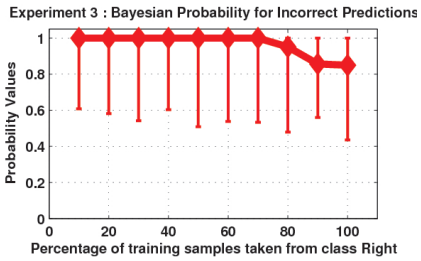
non-typical in their respective experiments, the 22.5° pose angle images seem to be more ambiguous. This is possibly because 90° pose angle images are, in a sense, 'extrema' in the label space, and hence, are less ambiguous to the system.

The results of Experiment 3 (Figure 2.23) once again reinforce our findings from Experiments 1 and 2. The BP and TRE values consistently stay close to 1 for incorrect predictions, irrespective of training bias. On the other hand, the confidence values in TCM show a very intuitive trend. For correct predictions, when a lesser number of face images with pose angle 'Right' are used for training, the framework very clearly shows a low confidence value, but gains confidence in its correct predictions as the relevant face images are increased (equivalent to gradually removing the training bias). Similarly, for incorrect predictions, although there are a couple of aberrations, the system stabilizes with a low confidence over time. However, in this case, the credibility values may be deceptive at times, and may need to be carefully interpreted in an appropriate manner.

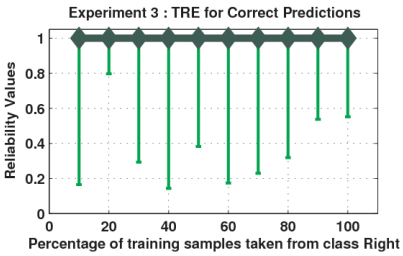
The results of Experiment 4 (Figure 2.24) establish the inference that the confidence measures obtained by the BP and TRE methods are not very informative about the prediction, and remain in the 0.9-1.0 range for all kinds of predictions across all training sample sizes. However, TCM once again shows a clear distinction in confidence and credibility values for correct and incorrect predictions. As mentioned before, it is worth re-iterating that this implication is extremely significant: *incorrect predictions can be filtered using a low confidence of the system*. The interesting inference, although, from this experiment is that the confidence of the system is not affected by variations in sample sizes in training. This may also be traced to the representativeness of the data points used in training, even if the number be small. This could be a potential direction for future work - to identify the most representative training points that can span a data universe.



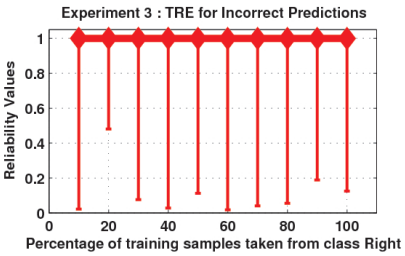
(a) BP for correct predictions



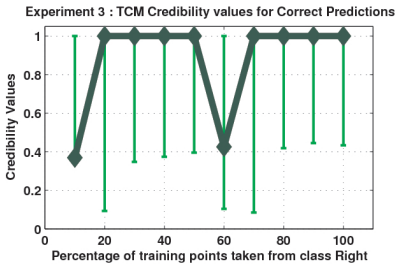
(b) BP for incorrect predictions



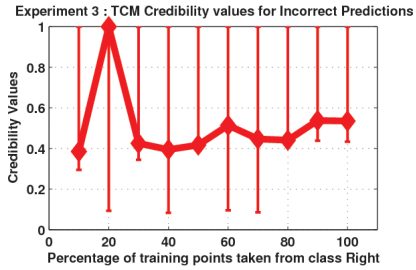
(c) TRE for correct predictions



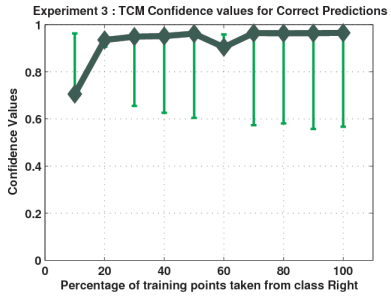
(d) TRE for incorrect predictions



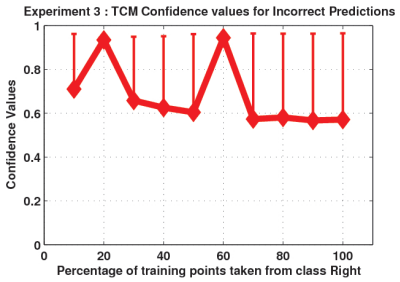
(e) TCM credibility for correct predictions



(f) TCM credibility for incorrect predictions



(g) TCM confidence for correct predictions



(h) TCM confidence for incorrect predictions

Figure 2.23: Results of Experiment 3: The x-axis denotes the increasing sample size (from 100 to 1000) used in consecutive steps, and y-axis the confidence values. The thick lines connect the median of the confidence values obtained across the test data points, while the thin lines along the vertical axis show the range of confidence values obtained at each sample size used for training

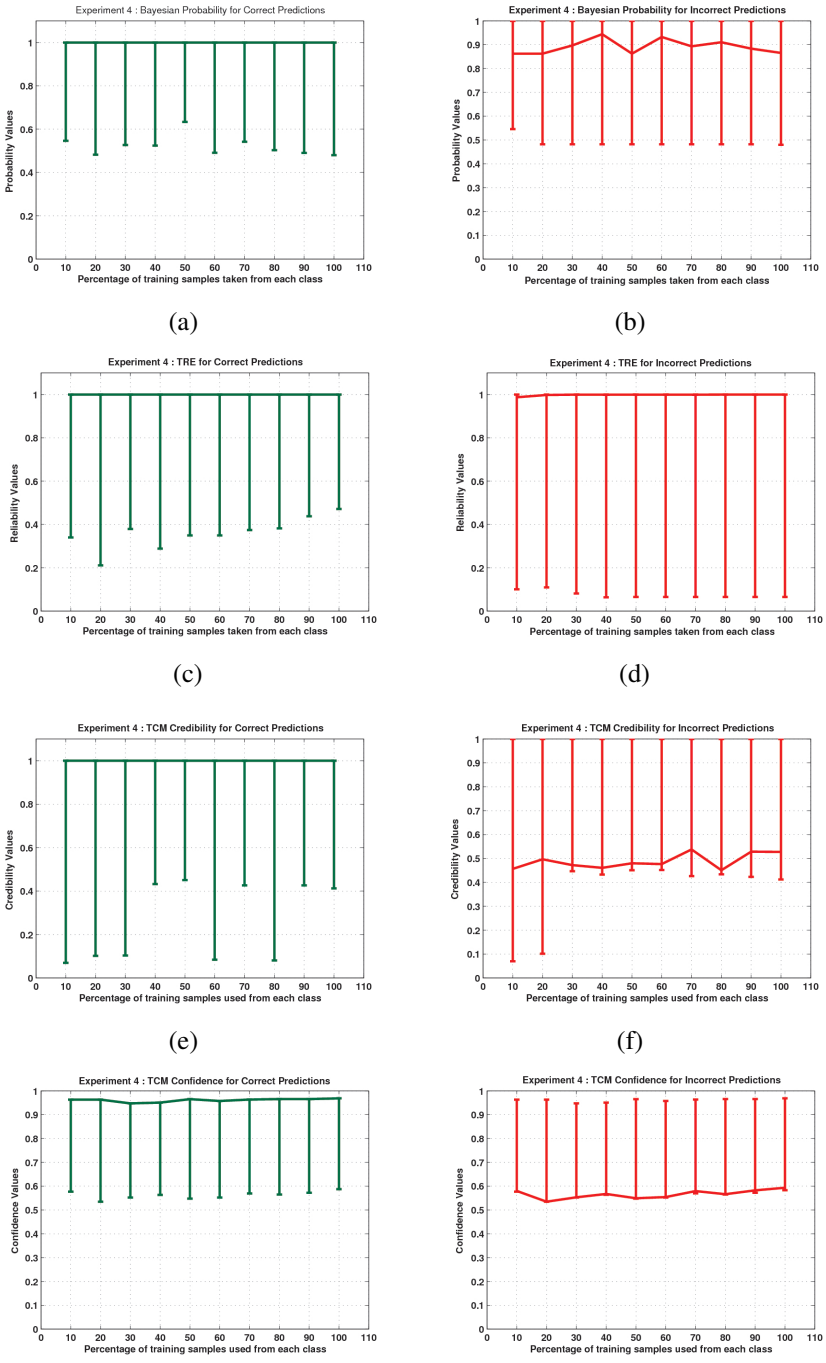


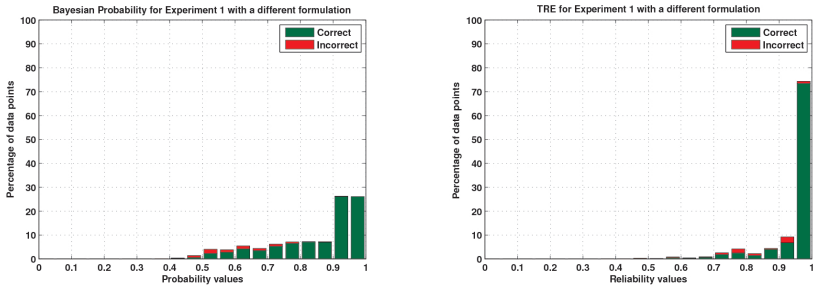
Figure 2.24: Results of Experiment 4

Inferences from the Study

This study provided an empirical validation of applying the CP (or TCM) framework for estimation of confidence in machine learning algorithms. At the outset, the CP framework seemed to address statistical issues associated with learning algorithms fairly well, in comparison to other popular methods. While not exhaustive, this study provided interesting insights on how the framework could be used in analyzing statistical causes of uncertainty. However, more work needs to be done to make such approaches practically viable.

To understand the poor results obtained consistently from the BP and the TRE methods, the formulation of the Adaboost.M2 algorithm with the k-NN classifiers was modified. In the initial formulation, the probability distribution on the output labels from the final boosted hypothesis was used as the probability value for the BP and TRE methods in the experiments. An additional study was performed by integrating the probabilities obtained in each of the iterations from the k-NN classifier (the probability distributions on the class labels from each iteration were added using the weights obtained from the Adaboost algorithm). In other words, the error margins (as defined in Table 2.4) from the k-NN output values in each of the boosting iterations have been incorporated into the frameworks. This however does not affect the current formulation of the TCM framework. The results of Experiment 1 with the new formulation for the BP and the TRE methods are shown in Figure 2.25.

As evident in Figure 2.25, the performance of the BP method improved, considering that the incorrect results have confidence values that are lower than in Figure 2.21. However, the confidence values for correct predictions seems to have fallen too. On the other end, although the reliability values obtained from TRE exhibit



(a) (b)

Figure 2.25: Results of Experiment 1 with a modified formulation for the BP and TRE methods

improved results (when compared to Figure 2.21), there is a significant chunk of incorrect predictions with a high reliability value. This certainly indicates that other formulations of the BP and the TRE methods may have performed better than the results presented in this study. This also indicates that the BP and the TRE methods rely on error margins to a significant extent(as defined in Table 2.4 to provide a value of confidence.

2.4 Summary

In this chapter, the background of this work, from both theory and application perspectives, were presented. The datasets used in this work were described, and tested for the validity of the assumptions for the CP framework. An empirical study of the performance of the CP framework, in comparison to two other frameworks, was performed, and the study showed significant promise in the framework's possibilities to provide a reliable measure of confidence under various conditions.

Chapter 3

EFFICIENCY MAXIMIZATION IN CONFORMAL PREDICTORS FOR CLASSIFICATION

As described in Chapter 2, the Conformal Predictions framework is a recent development in machine learning to associate reliable measures of confidence with results in classification and regression. While the formulation of the framework guarantees validity, the CP framework has another property called *efficiency* [38]. The efficiency of conformal predictors is measured by the size of the predicted set of labels (or the width of the predicted region in case of regression) at a particular confidence level. It is essential that the predicted sets are as small (or narrow) as possible. The efficiency of the framework depends greatly on the choice of the classifier and corresponding parameters such as kernel functions or distance metrics. While the CP framework has extensive potential to be useful in several applications, the lack of efficiency can limit its usability in real-world problems. For example, in a classification setting, the CP framework can output a prediction set containing all possible class labels for a given test point. Evidently, while this output is valid, such a result is not practically useful. Hence, it is essential that the CP framework, when applied in real-world problems, is both valid and efficient. This is the objective of this contribution of this dissertation. This chapter presents the methodology and results for efficiency maximization in the classification context, while Chapter 4 presents these details for the regression setting. The proposed methodology in this chapter is validated on the Cardiac Patient dataset as part of the Risk Prediction in Cardiac Decision Support application described in Chapter 2 (Section 2.2). Without any loss in generality, we describe the motivation and methodology in this

work assuming a binary classification problem for convenience of explanation and understanding.

3.1 Cardiac Decision Support: Background

As stated in Chapter 2, the objective of the work in this application domain is to predict the risk of complications following a coronary Drug Eluting Stent procedure (DES), using real-world patient data provided by Advanced Cardiac Specialists, a cardiology practice based in Arizona, USA. The dataset used to build the predictive model was described in Section 2.2.

Existing models in this scope (such as the Boston Scientific DES Thrombosis score [109]) have largely been rule-based and derived from correlation analysis. For example, in the DES Thrombosis score [109] approach, a set of patient attributes (between 5-10) are selected as correlated to the outcomes being studied. Suitable thresholds are identified for each of these attributes, and the predictive model is based on rules between these attribute-threshold pairs. A detailed listing of existing related models is presented in Table 3.1. The validity of such statistical models to *specific patient cases* is questionable. For example, age ≥ 65 is used as a common patient attribute in such models, and this may be set to zero for a patient with age 64. This increases the possibility of incidence of false positives and false negatives in the predictions, thereby limiting the scope of their applicability. Predictive models based on machine learning techniques have the ability to consider each patient as a unique entity, and predict outcomes for a particular patient case in question, unlike statistical models. Such a model has not been built for studying the risk of complications following a DES procedure, and this work is the first of its kind for this problem.

As mentioned earlier, the predictive model proposed in this work helps to strat-

MODEL NAME	SOURCE OBJECTIVE
EuroSCORE [110]	Assesses the European System for Cardiac Operative Risk Evaluation (EuroSCORE) validity to predict in-hospital mortality after PCI
Boston Scientific DES Thrombosis Risk Score [109]	A clinically useful patient risk score that predicts the incidence of stent thrombosis
Mayo Risk Score [111]	Identifies clinical and angiographic risk factors associated with complications of all kinds of PCI procedures
American College of Cardiology-National Cardiovascular Data Registry [112]	A risk adjustment model for in-hospital mortality following PCI procedures using data from a large, multi-center registry
Brigham and Women's Hospital [113]	Simplified risk score models for predicting the risk of major in-hospital complications after PCI in the era of widespread stenting and use of glycoprotein IIb/IIIa antagonists
University of Michigan Consortium [114]	Bedside prediction of prognosis for individual patients for PCI mortality.
Northern New England Cooperative Group [115]	Identifies risk factors associated with in-hospital mortality among patients undergoing PCIs.
Cleveland Clinic Foundation Multi-Center [116]	Establishes a relation between physician caseload and complication in PCI.
New York State [117]	Assesses the relationship between annual hospital volume and annual cardiologist volume for percutaneous transluminal coronary angioplasty (PTCA) and 2 outcomes of PTCA (in-hospital mortality and same-stay coronary artery bypass graft [CABG] surgery)
New York State [118]	Identifies significant independent risk factors for major percutaneous transluminal coronary angioplasty outcomes

Table 3.1: Existing models for risk prediction after a Percutaneous Coronary Intervention/Drug Eluting Stent procedure

ify the risk for a specific patient for post-DES complications, and thereby stratify patient populations according to healthcare requirements. In medical diagnosis/prognosis, it is not only important to provide a prediction, but also equally (or more) important to associate a measure of confidence with the prediction. In this work, the CP framework has been used to achieve this purpose. The CP framework ensures that the frequency of errors made in the model are calibrated, and hence, a physician can set a suitable confidence level and obtain corresponding predictions. However, the efficiency, or the number of class labels in the predicted set, can vary based on the choice of classifier or the choice of parameters. Maximal efficiency, along with validity, is critical in a risk-sensitive application for practical usability, and this is the motivation of this work, as described below.

3.2 Motivation: Why Maximize Efficiency

The output of the CP framework for a classifier such as k-Nearest Neighbors is a set of class labels, Γ^{ϵ}, as described in Chapter 2. In a binary classification problem, the output set predicted by the CP framework can contain zero, one or two (both) class labels. When the output set contains only one class label and this class label is correct (given the ground truth), this could be considered as the ideal solution. If the output contains only one class label which is however incorrect, this is termed as an *error*. If the output set contains zero predictions, we call that an *empty prediction*, which is also counted towards an *error* since this solution will not provide the user with the correct class label. If the output set contains both class labels, it is termed a *multiple prediction*, which however is not an error since it always contains the correct solution. For example, given C_1 and C_2 as the class labels, and a test data point x_{n+1} whose ground truth is C_1, an output of Φ is an empty prediction and an error; an output of C_2 is an error; and an output of $\{C_1, C_2\}$ is a multiple prediction.

Maximizing *efficiency* in the CP framework for a classification setting would be equivalent to having the least possible number of class labels in the prediction set. In other words, for high efficiency, *the number of test data points for which the CP framework provides multiple predictions as output should be as low as possible, while maintaining the same number of errors*.

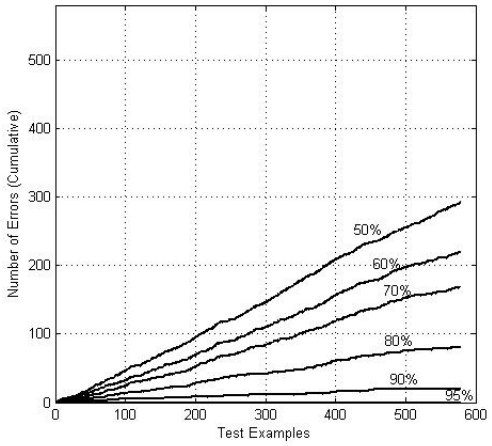
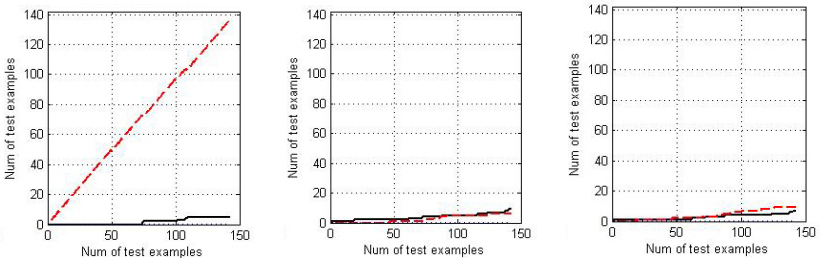


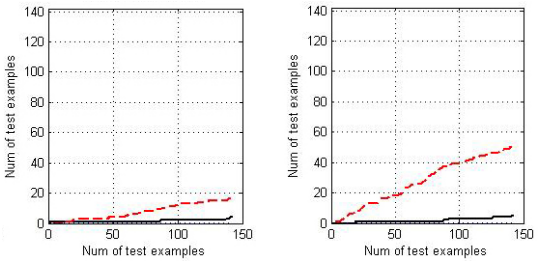
Figure 3.1: Illustration of the performance of the CP framework using the Cardiac Patient dataset. Note the validity of the framework, i.e. the errors are calibrated in each of the specified confidence levels. For example, at a 80% confidence level, the number of errors will always be lesser than 20% of the total number of test examples.

From the above discussion, it is evident that the performance of the CP framework can be summarized using two quantities: (i) number of errors, and (ii) number of multiple predictions. Since the CP framework guarantees validity [56], the number of errors will always remain bounded by $1-\varepsilon$ (as illustrated earlier in Figure 3.1). However, the efficiency of the framework lies in providing the maximum possible one-label prediction sets (at a given confidence level), since output sets with both labels in a binary classification problem do not provide any useful information to the end user. The efficiency can vary depending on the choice of a classifier, its

parameters or kernel functions. To illustrate this, Figure 3.2 presents the results of the CP framework for different classifiers and parameters at a user-specified 95% confidence level. As shown in the figure, efficiency varies with the choice of the classifier, while validity (number of errors shown with a black solid line) remains the same. The number of errors is always under 5% (100% - 95%, the specified confidence level), which is the CP framework's property. However, the number of multiple predictions varies for each classifier and corresponding parameter.



(a) SVM classifier, Gaussian kernel, spread=1 (b) k-NN classifier,k=5 (c) k-NN classifier,k=10



(d) k-NN classifier,k=20 (e) kernel k-NN classifier,k=5, Gaussian kernel, spread=5

Figure 3.2: Performance of CP framework on the Breast Cancer dataset from the UCI Machine Learning repository at the 95% confidence level for different classifiers and parameters. Note that the numbers on the axes are represented in a cumulative manner, as every test example is encountered. The black solid line denotes the number of errors, and the red dashed line denotes the number of multiple predictions

Similar experiments with the cardiac patient data were performed using a Sup-

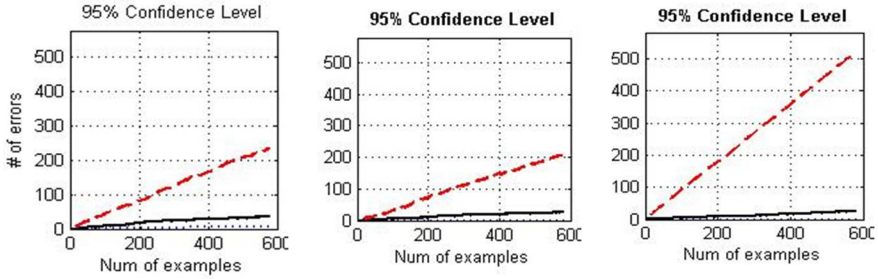
port Vector Machine (SVM) classifier with a polynomial kernel (which was found to give best results) and a neural network classifier, in addition to the k-NN classifier. 75% of the dataset was randomly selected for training, and the remaining as the testing subset. In case of SVM, the non-conformity measure for the CP framework was chosen to be:

$$\alpha_i^{y_p} = e^{-a \times d_i^m}$$

where d_i^m the distance of a given point i to the margin boundary of a class m. In addition, a back-propagation neural network was used on the same data with the non-conformity measure given by (as mentioned in Table 2.1 in Section 2.1):

$$\alpha_i^{y_p} = \frac{\sum_{y' \in Y : y' \neq y_p} o_{y'}}{o_{y_p} + \gamma}$$

The results obtained are presented in Figure 3.3. The black solid line denotes the number of errors, and it is evident that the number of errors is always under 5%, since the specified confidence threshold is 95%. However, the red dashed line denotes the number of multiple predictions in the binary classification problem. Note that this varies for each classifier, while the black solid line remains almost the same for all classifiers.



(a) SVM classifier (b) k-NN classifier (c) Back-propagation Neural Network

Figure 3.3: Performance of CP framework on the Cardiac Patient dataset

Note the line in red, which represents the number of multiple predictions. Although the same data was used and the frequency of errors was bounded by 5% at the presented 95% confidence level in both the cases, it is evident that the number of multiple predictions, which defines the efficiency, is significantly different (a 200% increase from SVM to NN) when a different classification algorithm is used. Since this is a binary classification problem, a prediction with multiple class labels is not practically useful, and it would be essential to minimize the number of multiple predictions (and thus maximize efficiency) to the extent that the choice of classifiers/non-conformity measures will allow. This motivates the need for a methodology that can minimize the number of multiple predictions, thus maximizing efficiency (while maintaining validity), given a particular classifier in the CP framework. This is the objective of this work.

The aforementioned limitation of the CP framework may act as a serious deterrent in its use in real-world applications, since it may not be an easy task to identify the correct parameters for a classifier that will provide the highest efficiency (or in other words, a practically useful conformal predictor). We propose an approach based on kernel learning to maximize efficiency in the CP framework. In particular, we learn an appropriate convex combination of kernel functions that can maximize efficiency, while maintaining validity. This methodology is validated using the k-NN classifier on datasets from the UCI Machine Learning repository [71], as well as the challenging Cardiac Patient dataset. The contributions of this work gain more value since there has been no earlier effort in this direction. We hope that this work will lead to adoption of the CP framework in real-world applications where there is a need for valid confidence measures. We now present the conceptual framework of the proposed methodology.

3.3 Conceptual Framework: Maximizing Efficiency in the CP Framework

As mentioned earlier, since this work is validated using a k-NN conformal predictor, we present the methodology for this classifier. However, the conceptual framework will remain similar for other classifiers, and can be extended conveniently. From the definition of a non-conformity measure for the k-NN classifier (Equation 2.6, Figure 2.1), it is evident that we would like the non-conformity measure for a data point that is assigned the correct class label (based on ground truth) to be as low as possible. Complementarily, we would like the non-conformity measure for a data point that is assigned an incorrect class label to be as high as possible. This will ensure that the p-value for the correct class label is very high, while the p-value for the incorrect class label is very low, thereby reducing the number of multiple predictions even at high confidence threshold levels. In order to achieve this, we would need to identify a kernel function, ϕ, such that for the projected data, $\phi(x)$:

- The margin between the classes is maximized
- The variance inside each of the classes is minimized

This is illustrated in Figure 3.4. Such a kernel feature space will ensure that the numerator of Equation 2.6 is low and the denominator is high, for a data point which is assigned the correct class label (and otherwise for an incorrect class label).

The first criterion - maximizing the margin - can be achieved using a Support Vector Machine(SVM)-based approach to kernel learning, as used in earlier work [119, 120, 121]. Similarly, the second criterion - minimizing intra-class variance - can be achieved by using a Linear Discriminant Analysis (LDA) [122] approach, i.e. by minimizing the denominator of the Fisher discriminant criterion, $w^T S_w w$, where S_w is the within-scatter matrix. Hence, the combination of these two

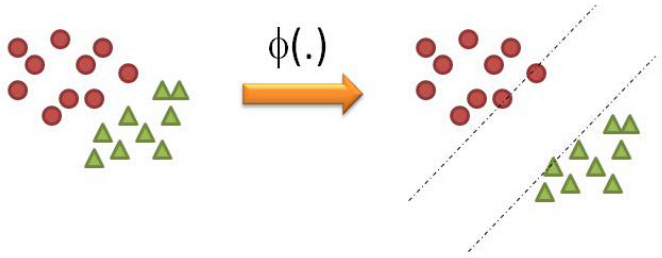


Figure 3.4: An illustration of an ideal kernel feature space for maximizing efficiency for a k-NN based conformal predictor

criteria can be used to learn a kernel function that can generate efficient conformal predictors.

The maximum-margin formulation, as in SVM [123], is given by (we assume a hard-margin formulation just for convenience of explanation. The implementations and results in this work use a soft-margin formulation):

$$\min \frac{1}{2} \|w\|^2$$

subject to $y_i(w^T x_i + b) \geq 1 \ \forall i = 1, 2, \ldots, n$. Combining the maximum-margin and minimum-variance criteria, the objective function can now be written as:

$$\begin{aligned} &\min \frac{1}{2} \|w\|^2 + w^T S_w w \\ &= \frac{1}{2} w^T w + w^T S_w w \\ &= \frac{1}{2} w^T (I + 2S_w) w \end{aligned}$$

More generally, this problem can now be written as:

$$\min \frac{1}{2} w^T (\lambda S_w + I) w \tag{3.1}$$

subject to $y_i(w^T x_i + b) \geq 1 \ \forall i = 1, 2, \ldots, n$, and where S_w is the within-class scatter matrix in Discriminant Analysis [10], and λ is a parameter that can be set empirically to balance the SVM and LDA components of the objective function. Note

that the within-class scatter matrix, S_w, is given by $\sum_{j=1}^{C}\sum_{i=1}^{N_j}(x_i - m_j)(x_i - m_j)^T$, where C is the number of class labels, N_j is the number of data points belonging to class j, and m_j is the mean vector of class j.

Now, substituting $\Lambda = \lambda S_w + I$, we get:

$$\min \frac{1}{2} w^T \Lambda w$$

subject to $y_i(w^T x_i + b) \geq 1 \; \forall i = 1, 2, \ldots, n$. The primal Lagrangian is then given by:

$$L(w, b, \alpha) = \frac{1}{2} w^T \Lambda w - \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + b) - 1)$$

Applying the KKT conditions and substituting back into the primal Lagrangian, we get the dual problem as:

$$\max L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j x_i^T \Lambda^{-1} x_j \tag{3.2}$$

subject to $\sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \geq 0 \; \forall i = 1, 2, \ldots, n$.

A similar formulation was used by Xiong and Cherkassky [124], but their approach was not used for Multiple Kernel Learning. However, they showed that the above formulation is equivalent to the following SVM formulation:

$$\min \frac{1}{2} \|\hat{w}\|^2$$

such that $y_i(\hat{w}^T \hat{x}_i + b) \geq 1 \; \forall i = 1, 2, \ldots, n$ where $\hat{w} = \Lambda^{1/2} w$ and $\hat{x}_i = \Lambda^{-1/2} x_i$ $\forall i = 1, 2, \ldots, n$. Evidently, this is the standard SVM formulation on the projected data points $\hat{x}_i$, and can be solved using existing SVM solving software. They also provided a method to compute $\Lambda^{1/2}$ and $\Lambda^{-1/2}$ using Singular Value Decomposition, which we have adopted in this work. Hence, the dual problem (Equation 3.2) to be solved can be rewritten as:

$$\max L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \hat{x}_i^T \hat{x}_j \tag{3.3}$$

subject to $\sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \geq 0 \; \forall i = 1, 2, \ldots, n$.

The optimization formulation to maximize efficiency in k-NN conformal predictors has thus been shown to be equivalent to a standard SVM problem with the projected data points.

However, this formulation has been derived for a linear SVM scenario. To extend this to a kernel learning setting, this modified SVM problem (using the projected data points) needs to be subsequently 'kernelized', and existing kernel learning methods that maximize the SVM margin can then be applied to this formulation. Details of the kernel learning methodology are presented in the following section. Note that while this formulation is one approach to maximizing efficiency, a more holistic solution will take into account the kernel function, even before projecting the data points onto a different space in the above derivation. Such an alternate formulation has been presented later in this chapter.

3.4 Kernel Learning for Efficiency Maximization

Kernel Learning: A Brief Review

The use of kernel methods in machine learning has grown immensely in the last decade. Kernel methods [125] [123] constitute a class of algorithms, where the data is mapped into a high-dimensional space where it is easier to find relations between data. These methods borrow their name from kernel functions, which enable them to operate in the feature space without computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the original data space. This is done using Mercer's theorem, which states that any continuous, symmetric, positive semi-definite kernel function $K(x, y)$ can be expressed as a dot product in a high-dimensional space[1].

[1] http://en.wikipedia.org/wiki/Kernel_trick

Kernel	Description
Polynomial	$k(x,x') =< x,x' >^d$
Radial Basis Function	$k(x,x') = f(d(x,x'))$, where d is a metric on X. Examples include Gaussian (please see below) and B-Splines.
Gaussian	$k(x,x') = \exp^{-\frac{\|x-\bar{x}\|^2}{2\sigma^2}}$
Sigmoid	$k(x,x') = \tanh(axx' + v)$

Table 3.2: Examples of kernel functions

More specifically, if the arguments to the kernel are in a measurable space X, and if the kernel is positive semi-definite, i.e. $\sum_{i,j} K(x_i,x_j)c_i c_j \geq 0$, then, for any finite subset $x_1,...,x_n$ of X and any real numbers $c_1,...,c_n$, then there exists a function $\varphi(x)$, whose range is in an inner product space of possibly high dimension, such that:

$$K(x,y) = \varphi(x) \cdot \varphi(y)$$

A few basic examples of kernel functions are listed in Table 3.2.

While kernel methods in the initial years relied on identifying optimal parameters by empirically minimizing classification error, there has been a recent growth in non-parametric approaches that can learn the kernel Gram matrix [121] or learn the weights of a convex combination of kernel functions [120] (commonly called Multiple Kernel Learning or MKL). MKL methods attempt to identify the optimal linear combination of kernel functions/matrices that maximize a performance measure, such as maximum margin classification error [121] or Fisher discriminant analysis [122]. Earlier work has shown such approaches to be promising in identifying the appropriate combination of kernel functions/matrices for improved performance. One of the earliest efforts in this regard was by Cristianini et al. [126], who proposed a methodology for kernel alignment. Given an unlabeled sample set $S = x_i : i = 1,\ldots,n$ and $x_i \in \Re^m$. kernels k_1 and k_2, then the inner product between

the kernel matrices is given by:

$$\langle K_1, K_2 \rangle_F = \sum_{i,j=1}^{n} k_1(x_i, x_j) k_2(x_i, x_j)$$

where K_i is denoted to be the kernel matrix based on the kernel k_i. Then, the alignment of the two kernels with respect to sample S is given by:

$$\hat{A}(S, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}}$$

One of the kernels can be based on the label vectors in the training set to ensure that the kernel matrix that is learnt, is aligned with the training data and the corresponding labels.

Since the work by Cristianini et al., there have been numerous approaches that have been proposed for kernel learning. A taxonomy of MKL algorithms presented by Gonen and Alpaydin [127] categorized such methods into fixed rules (fixed weights for each kernel function, for example), parametrized functions (linear and non-linear combinations of individual kernel functions), similarity-based methods (such as kernel alignment), boosting methods and Bayesian methods. In this work, we adopt an approach based on parametrized functions (in particular, a convex combination of kernel functions) that can maximize the efficiency of a classifier in the CP framework. The details of the proposed methodology are presented below.

Learning a Kernel to Maximize Efficiency

Similar to Lanckriet's formulation [121], Equation 3.3 can now be rewritten in the context of a Multiple Kernel Learning Problem as:

$$\min_{p \in P} \max_{\alpha \in Q} f(\mathbf{p}, \alpha) = \alpha^T \mathbf{e} - \frac{1}{2} (\alpha \circ \mathbf{y})^T (\sum_{i=1}^{m} p_i K_i)(\alpha \circ \mathbf{y})$$

where $P = \left\{\mathbf{p} \in \mathbb{R}^m : \mathbf{p}^T\mathbf{e} = 1, 0 \leq \mathbf{p} \leq 1\right\}$ denotes the set of kernel weights, Q is the set of SVM dual variables such that:

$$Q = \left\{\alpha \in \mathbb{R}^n : \alpha^T\mathbf{y} = 0, \alpha \geq 0\right\}$$

$\mathbf{e}$ is a vector all ones, $\{K_i\}, i = 1, 2, \ldots, m$ is a group of base kernel matrices that are defined on the projected data, $\hat{x}_i$, and $\circ$ denotes the vector dot product.

Several ways of solving this optimization problem have been proposed in the past. The MKL problem was first formulated as a semi-definite programming problem [121]. More recent approaches include Quadratically Constrained Quadratic Programming [122], Sequential Minimal Optimization [128], Semi-Infinite Linear Programming [129] and Subgradient Descent [120]. Despite the success of many of these methods, each of them has its own limitations. For example, as pointed out in [119], the Subgradient Descent method uses the gradient of only the current solution in its iterative computations, and the Semi-Infinite Linear Programming method does not regularize the approximate solution obtained from the cutting plane model. Addressing these limitations, Xu et al. [119] proposed a method based on the level method [130], which is generally used to solve non-smooth optimization problems. Here, the level method is extended to address min-max optimizations (convex-concave optimization, to be precise) and thus applied to MKL. Their results demonstrate that this method is more efficient than other existing methods, and hence, is used in this work. The algorithm for this Extended Level method for MKL is summarized below. More details of this method can be found in [119].

In order to learn a convex combination of kernel functions, the base kernels can be created in different ways: by using different kernel functions, or just a single kernel function with different parameter values or subsets of features. In this work, we combined both these approaches and allowed the algorithm to select the most ap-

Algorithm 3 The Extended Level Method for Multiple Kernel Learning

Require: Individual kernel Gram matrices, $\{\mathbf{K}_i, i = 1,\ldots,m\}$, kernel weights vector $\mathbf{p}$, $\mathbf{e}$ is a vector of all ones, termination threshold ε

1: Initialize $\mathbf{p}^0 = \frac{\mathbf{e}}{m}$ and i=0
2: **repeat**
3: Solve the dual problem of SVM with $\mathbf{K} = \sum_{j=1}^{m} p_j^i \mathbf{K}_j$ to obtain the optimal solution α^i
4: Construct the cutting plane model, $g^i(\mathbf{p}) = \max_{1\leq j\leq i} f(\mathbf{p}, \alpha^j)$, where f is defined in Equation 3.4
5: Calculate the lower bound $\underline{f}^i = \min_{\mathbf{p}\in P} g^i(\mathbf{p})$ and the upper bound $\overline{f}^i = \min_{1\leq j\leq i} f(\mathbf{p}^j, \alpha^j)$, and the gap $\Delta^i = \overline{f}^i - \underline{f}^i$
6: Compute the projection of $\mathbf{p}^i$ onto the level set L^i by solving the optimization problem $\mathbf{p}^{i+1} = \arg\min_{\mathbf{p}} \left\{ \left\|\mathbf{p} - \mathbf{p}^i\right\|_2^2 : \mathbf{p} \in P, f(\mathbf{p}, \alpha^j) \leq l^i, j = 1,\ldots,i \right\}$
7: Update $i = i + 1$
8: **until** $\Delta^i \leq \varepsilon$

propriate kernel functions and parameter values that can maximize the efficiency of the k-NN conformal predictor. The validation of the proposed method on different datasets is presented in the next section.

3.5 Experiments and Results

Data Setup

To study the performance and generalizability of the proposed method, we carried out experiments on three binary datasets (with different number of dimensions and instances): 2 datasets from the UCI Machine Learning repository, and the challenging Cardiac Patient dataset. We focused on datasets from the healthcare domain, since reliable confidence measures are extremely valuable for machine learning algorithms to be successfully applied in this domain. The datasets and their details are listed in Table 6.1. 75% of each of the datasets was randomly permuted (to meet the exchangeability requirements of the CP framework) and used for training, while the remaining portion of the dataset was used for testing. Further, the experiment

was repeated 5 different times to remove any randomness bias.

Dataset	Size of dataset	Dimensions
SPECT	267	22
Breast Cancer	569	30
Cardiac Patient	2312	165

Table 3.3: Datasets used in our experiments

Details about the SPECT and Breast Cancer datasets can be found on the UCI repository website [71]. The Cardiac Patient dataset has been described earlier in Section 2.2. This dataset contains the profiles of 2312 patient cases (with a set of 165 attributes each) who had a Drug Eluting Stent procedure performed during the period 2003-07, and who had followed up with the facility during the 12 months following the procedure. This is formulated as a binary classification problem which predicts the onset of complications, or otherwise. More details of the dataset can also be found in [131].

Experimental Results

In the CP framework, the validity property is always satisfied by definition, i.e. the number of errors are always bounded by the confidence threshold. This was also empirically confirmed in our work, as shown in Figures 3.1, 3.2 and 3.3. In this section, we focus on studying the results related to the efficiency of the CP framework. The proposed MKL approach was compared against the plain k-NN classifier (with different values of k) and kernel k-NN classifier with varying kernel functions and parameters.

Tables 3.4, 3.5 and 3.6 present these results for the SPECT, Breast Cancer and Cardiac Patient datasets respectively. Note that in each of these tables, the best representative results were selected and presented for each of the considered classifiers, since it was not possible to present the results obtained with all possible

parameters. Similar results were obtained for other combinations of kernel parameters that have not been included in the table too, but the results in the tables were the most representative of the observed trends. Also, while the number of errors were not mentioned in the tables (for clarity of presentation), it was verified that the validity property continued to hold good for the CP framework on applying the proposed kernel learning method. The number of multiple predictions is the number of test data points that the k-NN conformal predictor provided both class labels as an output. *A lower number of multiple predictions at all possible confidence levels is most desirable*.

Classifier	Parameters	Number of Multiple Predictions at Confidence Level (Total: 66 test points)				
		70%	80%	90%	95%	99%
k-NN	k=3	0	0	9	33	56
k-NN	k=10	0	0	10	34	66
kernel k-NN	k=3, Gaussian kernel, Spread=100	0	0	8	35	54
kernel k-NN	k=10, Gaussian kernel, Spread=100	0	0	11	26	66
kernel k-NN	k=3, Polynomial kernel, Degree=2	0	0	8	31	54
kernel k-NN	k=3, Polynomial kernel, Degree=3	0	2	16	37	56
Proposed MKL	k=5, Mixture of Polynomial kernels	0	0	1	20	59
Proposed MKL	k=5, Mixture of Gaussian kernels	0	0	2	17	60
Proposed MKL	k=5, Mixture of Polynomial and Gaussian kernels	0	0	0	23	46

Table 3.4: Results obtained on the SPECT Heart dataset. Note that the number of multiple predictions are clearly the least when using the proposed MKL approach, even at high confidence levels

All the tables unanimously validate that the results obtained with the proposed

Classifier	Parameters	Number of Multiple Predictions at Confidence Level (Total: 142 test points)				
		70%	80%	90%	95%	99%
k-NN	k=3	0	0	0	8	42
k-NN	k=5	0	0	0	6	60
kernel k-NN	k=3, Gaussian kernel, Spread=25	0	1	3	31	73
kernel k-NN	k=5, Gaussian kernel, Spread=25	0	0	2	28	72
kernel k-NN	k=3, Polynomial kernel, Degree=2	0	0	0	8	52
kernel k-NN	k=5, Polynomial kernel, Degree=3	0	0	0	7	61
Proposed MKL	k=5, Mixture of Polynomial kernels	0	0	0	0	43
Proposed MKL	k=5, Mixture of Gaussian kernels	0	0	0	0	37
Proposed MKL	k=5, Mixture of Polynomial and Gaussian kernels	0	0	0	0	36

Table 3.5: Results obtained on the Breast Cancer dataset. Note that the number of multiple predictions are clearly the least when using the proposed MKL approach, even at high confidence levels

MKL approach for efficiency maximization are significantly better than the best possible results obtained with the other classifiers (which were themselves obtained after long trials of varying parameter values). It can be observed that the number of multiple predictions are very low at lower confidence levels. This is because the CP framework allows a higher number of errors at lower confidence levels, thereby reducing the number of multiple predictions. Hence, it is rather most desirable to obtain low number of multiple predictions at very high confidence levels. Note that when the number of multiple predictions is high, the classifier is providing results with both class labels, thereby serving no purpose to the physician in prognosing or diagnosing the patient. The proposed approach reduces this number significantly to

Classifier	Parameters	Number of Multiple Predictions at Confidence Level (Total: 578 test points)				
		70%	80%	90%	95%	99%
k-NN	*k*=3	0	0	1	181	522
k-NN	*k*=15	0	0	0	193	500
kernel *k*-NN	*k*=3, Gaussian kernel, Spread=1	0	0	5	190	496
kernel *k*-NN	*k*=3, Gaussian kernel, Spread=10	0	0	1	184	512
kernel *k*-NN	*k*=3, Polynomial kernel, Degree=2	0	0	1	193	519
kernel *k*-NN	*k*=3, Polynomial kernel, Degree=3	0	0	2	176	517
Proposed MKL	*k*=3,5,10; Mixture of Polynomial kernels	0	0	0	141	461
Proposed MKL	*k*=3,5,10; Mixture of Gaussian kernels	0	0	0	137	470
Proposed MKL	*k*=3,5,10; Mixture of Polynomial and Gaussian kernels	0	0	0	136	462

Table 3.6: Results obtained on the Cardiac Patient dataset. Note that the number of multiple predictions are clearly the least when using the proposed MKL approach, even at high confidence levels.

provide more useful results to the end user.

3.6 Discussion

Additional Results

Since the kernel learning formulation described in the previous section can be viewed as related to the objectives of the Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) classifiers, we conducted further experiments with related combinations of classifiers to compare with the performance of the proposed methodology. The methods that were considered in this study included:

1. **LDA-*k*NN:** Linear Discriminant Analysis (LDA) [10] was used to compute

the projections of data points, and the CP framework based on the k-NN classifier was subsequently applied to the projected data points.

2. **KLDA-kNN:** Kernelized Linear Discriminant Analysis (LDA) (with Gaussian and polynomial kernels using parameters studied in the previous section) was used to compute the projections of data points, and the CP framework based on the k-NN classifier was subsequently applied to the projected data points.

3. **DMKL-kNN:** The Discriminant Multiple Kernel Learning approach proposed by Ye et al. [122] was used to learn the kernel, following which the k-NN classifier was applied. This method formulates the MKL problem based on LDA, and uses Quadratically Constrained Quadratic Programming (QCQP) to solve the problem. Once again, the Gaussian and polynomial kernel functions with parameters used in the previous section were used in this MKL procedure.

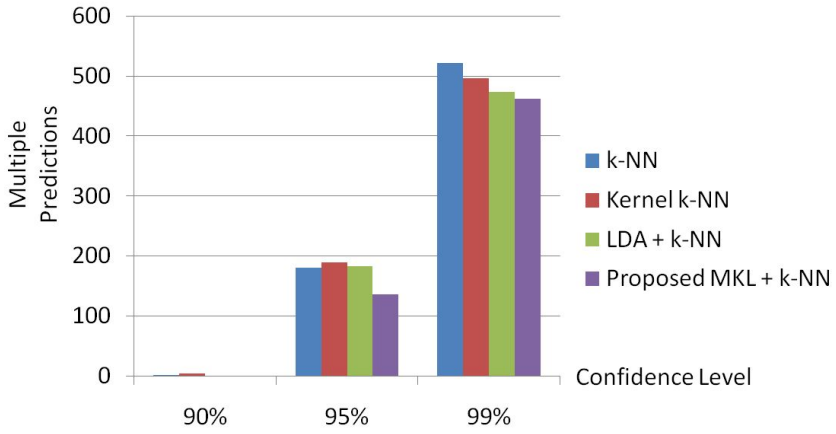


Figure 3.5: Summary of results showing the number of multiple predictions on the Cardiac Patient dataset using various methods including the proposed MKL method. Note that kernel LDA + kNN also provided results matching the proposed framework

The most representative results for these experiments are shown in Tables 3.7, 3.8 and 3.9. A summary of the results for the Cardiac Patient dataset is also presented in Figure 3.5. Similar to the previous results, only the number of multiple predictions is presented for each method for clarity of presentation (since the validity property was found to hold good). The results of the proposed method were reproduced here for ease of comparison. On observation, these results led to the following inferences:

- Firstly, among the MKL (kernel learning) methods, the proposed method provided the best results. This validates our approach, and places merit in using this method for maximizing efficiency in the CP framework.
- Interestingly, KLDA-kNN (kernel LDA followed by k-NN with the CP framework) showed better or equivalent results, when compared to the proposed method. From one perspective, this result only indicates that the proposed MKL approach can be applied to learn a kernel that is used with LDA to compute projections of data, instead of using the learnt kernel directly with k-NN. In other words, other classifiers may have given better results than k-NN. But considering that this work was carried out as a proof-of-concept with k-NN as the classifier, a similar reasoning could be applied with LDA to derive an appropriate MKL procedure.

Alternate Formulation

The kernel learning formulation proposed in Section 3.4 derived an objective function to be optimized, and subsequently, the kernel version of the objective function was used to learn a convex combination of kernel Gram matrices using standard Multiple Kernel Learning methods. While this method demonstrated satisfying results, it is possible to formulate the objective function for optimization, by including

Classifier	Parameters	Number of Multiple Predictions at Confidence Level (Total: 66 test points)				
		70%	80%	90%	95%	99%
LDA-kNN	k=5	0	0	9	42	56
KLDA-kNN	k=25,Gaussian kernel,Spread=0.1	0	0	1	10	13
KLDA-kNN	k=3,Polynomial kernel,Degree=1	0	0	1	16	56
DMKL-kNN	k=3,Mixture of Polynomial kernels	0	4	17	37	56
DMKL-kNN	k=3,Mixture of Gaussian kernels	0	1	9	38	56
DMKL-kNN	k=3,Mixture of Polynomial and Gaussian kernels	0	3	17	37	56
Proposed MKL	k=5, Mixture of Polynomial kernels	0	0	1	20	59
Proposed MKL	k=5, Mixture of Gaussian kernels	0	0	2	17	60
Proposed MKL	k=5, Mixture of Polynomial and Gaussian kernels	0	0	0	23	46

Table 3.7: Additional results on the SPECT Heart dataset

the kernel function from the very beginning, instead of 'kernelizing' the objective function at the end. One such possibility is presented below.

We begin this discussion from Equation 3.1 in the formulation presented in Section 3.4. The 2 criteria for maximizing efficiency in the CP framework using k-NN can be satisfied through the following optimization problem.

$$\min_{w} \frac{1}{2} w^T (\lambda S_w + I) w \tag{3.4}$$

subject to $y_i(w^T x_i + b) \geq 1 \ \forall i = 1, 2, \ldots, n$, and where $S_w = \sum_i \sum_x (x - m_i)^2$, the within-scatter matrix in Discriminant Analysis. λ is a parameter that can be set empirically to balance the SVM and LDA components of the objective function.

Classifier	Parameters	Number of Multiple Predictions at Confidence Level (Total: 142 test points)				
		70%	80%	90%	95%	99%
LDA-kNN	k=25	0	0	0	11	53
KLDA-kNN	k=5,Gaussian kernel,Spread=100	0	0	0	1	21
KLDA-kNN	k=5,Polynomial kernel,Degree=2	0	1	33	46	103
DMKL-kNN	k=5,Mixture of Polynomial kernels	0	0	0	7	61
DMKL-kNN	k=5,Mixture of Gaussian kernels	0	0	0	11	46
DMKL-kNN	k=5,Mixture of Polynomial and Gaussian kernels	0	0	0	7	61
Proposed MKL	k=5, Mixture of Polynomial kernels	0	0	0	0	43
Proposed MKL	k=5, Mixture of Gaussian kernels	0	0	0	0	37
Proposed MKL	k=5, Mixture of Polynomial and Gaussian kernels	0	0	0	0	36

Table 3.8: Additional results on the Breast Cancer dataset

The above objective function can be written as:

$$\min_{w} \left[\frac{\lambda}{2} w^T S_w w + \frac{1}{2} w^T w \right]$$

From the Kernel Fisher Discriminants work of Mika [132], we have that $w^T S_w w$ can be written in kernel space as:

$$\alpha^T N \alpha$$

where $N = KDK^T$, K is the kernel Gram matrix, $D = I - V_1 V_1^T - V_2 V_2^T$, I is the identity matrix, V_j is the vector with element $(V_j)_i) = \frac{1}{\sqrt{M_j}}$ if the example i belongs to class j and zero otherwise, M_j is the number of data points belonging to class j.

Classifier	Parameters	Number of Multiple Predictions at Confidence Level (Total: 578 test points)				
		70%	80%	90%	95%	99%
LDA-*k*NN	*k*=20	0	0	0	183	474
KLDA-*k*NN	*k*=25,Gaussian kernel,Spread=5	0	0	0	136	443
KLDA-*k*NN	*k*=20,Polynomial kernel,Degree=1	0	0	0	136	471
DMKL-*k*NN	*k*=3,Mixture of Gaussian kernels	0	0	5	190	496
Proposed MKL	*k*=3,5,10; Mixture of Polynomial kernels	0	0	0	141	461
Proposed MKL	*k*=3,5,10; Mixture of Gaussian kernels	0	0	0	137	470
Proposed MKL	*k*=3,5,10; Mixture of Polynomial and Gaussian kernels	0	0	0	136	462

Table 3.9: Additional results on the Cardiac Patient dataset

Similarly, in kernel space, using the Representer Theorem:

$$w^T w = (\sum_i \alpha_i \phi(x_i))^T (\sum_j \alpha_j \phi(x_j))$$

Note that αs above will be the same as the αs in the expression for $w^T S_w w$, since the corresponding derivation for the latter in [132] begins with the same expansion using the Representer Theorem. Continuing with the above expression, we get:

$$w^T w = \sum_i \sum_j \alpha_i^T \phi(x_i)^T \phi(x_j) \alpha_j = \sum_i \sum_j \alpha_i^T k(x_i, x_j) \alpha_j$$

$$= \alpha^T \mathbf{K} \alpha$$

Hence, the objective function in Equation 3.4 can now be rewritten as:

$$\min_\alpha \left[\frac{\lambda}{2} \alpha^T \mathbf{N} \alpha + \frac{1}{2} \alpha^T \mathbf{K} \alpha \right]$$

$$= \min_\alpha \frac{1}{2} \alpha^T (\lambda \mathbf{N} + \mathbf{K}) \alpha \tag{3.5}$$

Let $\beta = \lambda \mathbf{N} + \mathbf{K}$. Equation 3.5 can now be written as:

$$\min_{\alpha} \frac{1}{2} \alpha^T \beta \alpha \tag{3.6}$$

In a similar manner, the original constraint $y_i(wx_i + b) \geq 1$ can be written in the kernel space as:

$$y_i((\sum_i \alpha_j \phi(x_j))\phi(x_i) + b) \geq 1 \forall i = 1, \ldots, n \tag{3.7}$$

Let K_{x_i} denote the column of $\mathbf{K}$ corresponding to the i^{th} data instance. Then, Equation 3.7 can be rewritten as:

$$y_i(\alpha^T K_{x_i} + b) \geq 1 \forall i = 1, \ldots, n \tag{3.8}$$

The primal Lagrangian for the above objective function is then given by:

$$\mathbf{L}(\alpha, b, \gamma) = \frac{1}{2} \alpha^T \beta \alpha - \sum_i \gamma_i \left[y_i(\alpha^T K_{x_i} + b) - 1 \right]$$

where γ_is are the Lagrange multipliers. Using KKT conditions, we get:

$$\nabla \mathbf{L}_\alpha = 0 \Rightarrow \alpha = \beta^{-1} \sum_i \gamma_i y_i K_{x_i}$$

$$\nabla \mathbf{L}_b = 0 \Rightarrow \sum_i \gamma_i y_i = 0$$

Substituting this into the primal Lagrangian, we get the dual Lagrangian as:

$$\mathbf{L}(\gamma) = \sum_i \gamma_i - \frac{1}{2} \sum_i \sum_j y_i y_j \gamma_i \gamma_j K_{x_j}^T \beta^{-1} K_{x_i} \tag{3.9}$$

$$s.t. \sum_i \gamma_i y_i = 0, \gamma_i \geq 0 \forall i = 1, 2, \ldots, n$$

If the substitution $\hat{x_i} = \beta^{-1} K_{x_i}$ is used for all $i = 1, 2, \ldots, n$, the above formulation will be equivalent to a traditional linear SVM, and can be solved using traditional SVM solvers. However, it needs to be investigated if this substitution can be made in a straightforward manner, or if this can be solved further analytically. This will form a component of the future directions of this work.

3.7 Summary

The Conformal Predictions framework is a recent game-theoretic approach to compute reliable confidence measures across all kinds of machine learning algorithms. While it provides confidence measures that are valid in terms of the frequency of errors, this framework is not efficient enough to be practically useful in real-world applications. In this chapter, we proposed a new methodology to maximize the efficiency of the CP framework using Multiple Kernel Learning. We validated our approach using the k-NN classifier and a MKL method that maximized the margin and minimized intra-class variance in binary classification problems. This MKL problem was solved using the Extended Level Set method. While we validated our approach using k-NN, this methodology can be adapted to any other classifier, depending on the definition of the non-conformity measure for the classifier. The results that we obtained clearly showed the reduction in the number of multiple predictions, even at very high confidence levels, thus increasing the efficiency of the conformal predictor, while maintaining its validity. Such efficient conformal predictors can be of high practical value to end users. This was illustrated through the results we obtained with the Cardiac Patient dataset.

The results obtained in this work demonstrate great promise and corroborate the potential of applying kernel learning for maximizing efficiency in conformal predictors given a particular classifier. In future work, we intend to study the possibility of identifying a universal framework to maximize efficiency, irrespective of the classifier being used.

3.8 Related Contributions

In addition to the contributions based on the CP framework, there were other contributions that were made to address related problems in cardiac decision support

using the Cardiac Patient dataset. These related contributions have been briefly described below.

Synthetic Minority Oversampling Technique for Handling Imbalanced Data in Medical Domain: A significant challenge in this work, common to most clinical machine learning problems, was imbalanced data, and we implemented the Synthetic Minority Over-sampling Technique (SMOTE) to address this issue. Our results demonstrated the effectiveness of SMOTE in handling imbalanced data with a SVM approach. The developed predictive model provided an accuracy of 94% with a 0.97 AUC (Area under ROC curve), indicating high potential to be used as a decision support for management of patients following a DES procedure in real-world cardiac care facilities. More details of this work can be found in [78].

Patient Attribute Selection using Recursive Feature Elimination: In this work, we used SVM based Recursive Feature Elimination (SVM-RFE) methods to select patient attributes/features relevant to the etio-pathogenesis of complications following a drug eluting stent (DES) procedure. With a high dimensional feature space (165 features, in our case), and comparatively few patients, there is a high risk of 'over-fitting'. Also, for the model to be clinically relevant, the number of patient features need to be reduced to a manageable number, so that such an approach can be adopted in patient care. SVM-RFE selects subsets of patient features that have maximal influence on the risk of a complication. In our results, when compared with our initial model with all the 165 features, we obtained better performance of the classifiers with 75 top ranked patient features, a 50% reduction in the original dimensionality of the data space. There was a universal improvement in performance of all SVMs with different kernels and parameters. This method of feature ranking helps to determine the most informative patient features. Use of these relevant features improves the prediction of complications following a DES procedure.

More details of this work can be found in [133] and [134].

Clinically Relevant Ontology-Based Kernel Methods for Risk Prediction: Machine learning frameworks that are used to build clinical predictive models are founded on the concept of inter-data distance metrics. Algorithms in non-medical domains use distance metrics such as the Euclidean, Manhattan or the Mahanalobis distance. However, in clinical machine learning, a significant challenge is the inherent nature of data in the medical domain, since terms in the medical domain have strong interdependencies and hierarchical relationships. In this work, we developed a clinically relevant inter-patient kernel metric that is based on the patient data at hand, and the SNOMED medical ontology[2], which contains over a million medical concepts and is commonly used in healthcare. Using these knowledge-driven injected kernels resulted in the improved performance of risk classification over traditional kernels. From a broader perspective, this work revealed that the use of domain knowledge in predictive modeling enables the development of better models for clinical decision support. Although the current work is on a population of patients with DES, contributions of this work can be generalized to all clinical machine learning frameworks across other medical domains. More details of this work can be found in [134].

[2]http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

Chapter 4

EFFICIENCY MAXIMIZATION IN CONFORMAL PREDICTORS FOR REGRESSION

The regression formulation of the Conformal Predictions framework necessitates an approach that is different from that of the methodology proposed for classification in Chapter 3. This chapter presents the methodology and results for efficiency maximization in the regression context for the Head Pose Estimation application described in Chapter 2 (Section 2.2). The key idea of the proposed method in this chapter is to learn a suitable distance metric that can maximize the efficiency of the Conformal Prediction framework for ridge regression.

Many kinds of approaches have been adopted to solve the pose estimation problem in recent years. A broad subjective classification of these techniques based on earlier attempts [135] [136] [137] [138] is summarized in Table 4.1 [139]. Shape-based geometric and appearance-based methods have been the most popular approaches for many years. More recently, it has been shown that face images with varying poses can be visualized as lying on a smooth low-dimensional manifold, and this has opened up efforts to approach the problem from the perspectives of non-linear dimensionality reduction - more commonly called *manifold learning*. In this work, we propose a method for maximizing efficiency in ridge regression using manifold learning as a means of distance metric learning.

4.1 Motivation: Why Maximize Efficiency in Regression

When the algorithm for the CP framework in the regression context is applied (as outlined in Algorithm 4 in Section 2.1), the result is presented as a union of intervals each of whose p-value is greater than the specified confidence level. Hence, a

natural measure of efficiency in regression problems is the length of the final predicted interval. Vovk et al. [38] also used the median length of the convex closures of prediction sets as a measure of efficiency of a sequence of predictions (when many non-intersecting neighborhoods form the prediction region in regression). In this work with head pose estimation, preliminary experiments suggested that the prediction regions were most often just a single interval (and not a union of non-intersecting neighborhoods). Hence, the mean width of the predicted intervals for a set of test data instances is used as a measure of efficiency, as in [66].

When the CP framework algorithm is applied to the FacePix dataset for head pose estimation (as described in Section 2.2), the frequency of errors and the mean width of the predicted regions are presented in Table 4.2. Note that the the results presented in the table were the best and most representative results obtained from many empirical trials with different combinations of parameters. Details of these experiments are presented later in this chapter. As Table 4.2 suggests, the percentage of errors are always calibrated with respect to the specified confidence level. This is a very useful result, and can be practically very useful in real-world multimedia pattern recognition problems. For results on the mean widths, it should be mentioned that the range of head pose values is $[-90^{\circ}, +90^{\circ}]$, i.e. a total range of 181. As seen in the table, the mean width of the prediction region at almost all of

Shape-based geometric methods	[140] [141] [137] [142] [143]
Model-based methods	[144] [145] [146] [80]
Appearance-based methods	[147] [148] [149] [150] [151] [152]
Template matching methods	[153] [154]
Dimensionality reduction based approaches	[136] [155] [156] [157] [158] [135] [138]

Table 4.1: Classification of methods for pose estimation

the confidence levels encompasses the entire range (rather, is larger in some cases). Such a prediction region is practically of very less value in decision-making. Narrowing the prediction interval while maintaining validity, a task called maximizing efficiency in this work, is the objective in this chapter.

Feature	Percentage of Errors at Confidence Level				
	70%	80%	90%	95%	99%
Grayscale Pixel Intensity	30.04	19.19	8.61	4.32	1.21
Laplacian of Gaussian	30.04	20.07	9.96	5.02	1.03

Feature	Mean Width of Prediction Region at Confidence Level				
	70%	80%	90%	95%	99%
Grayscale Pixel Intensity	184.64	212.64	235.43	246.98	262.93
Laplacian of Gaussian	128	144	164	172	180

Table 4.2: Results of the CP framework for regression on the FacePix dataset for head pose estimation

4.2 Conceptual Framework: Maximizing Efficiency in the Regression Setting

The Conformal Predictions algorithm for ridge regression used in this work was presented in Algorithm 4 in Section 2.1. A detailed derivation of this method can be found in [66]. In order to effectively present the conceptual framework of maximizing efficiency in this section, we reproduce the algorithm here for ease of understanding.

In this algorithm, the final predicted regions are specified by the $\hat{y}_i$ values, which is a sorted array of all the u_i and v_i values. u_i and v_i are determined for each training data point individually. Hence, in order to keep the final predicted interval as narrow as possible, one possible solution would be to make u_i and v_i as close to each other as possible for all training data. Note that there may be other ways of achieving this objective, and this approach is one possibility.

Algorithm 4 Conformal Predictors for Regression

Require: Training set $T = \{(x_1, y_1), ..., (x_n, y_n)\}$, new example x_{n+1}, confidence level r, $X = x_1, x_2, \ldots, x_{n+1}$

1: Calculate $C = I - X(X'X + \alpha I)^{-1}X'$ (for ridge regression).
2: Let $A = C(y_1, y_2, \ldots, y_n, 0)' = (a_1, a_2, \ldots, a_{n+1})$
3: Let $B = C(0, 0, \ldots, 0, 1)' = (b_1, b_2, \ldots, b_{n+1})$
4: **for** $i = 1$ to $n+1$, do **do**
5: Calculate u_i and v_i.
If $b_i \neq b_{n+1}$, then $u_i = min(\frac{a_i - a_{n+1}}{b_{n+1} - b_i}, \frac{-(a_i + a_{n+1})}{b_{n+1} + b_i})$; $v_i = max(\frac{a_i - a_{n+1}}{b_{n+1} - b_i}, \frac{-(a_i + a_{n+1})}{b_{n+1} + b_i})$
If $b_i = b_{n+1}$, then $u_i = v_i = \frac{-(a_i + a_{n+1})}{2b_i}$.
6: **end for**
7: **for** $i = 1$ to $n+1$, do **do**
8: Compute S_i according to Equation 2.10 below.
9: **end for**
10: Sort $(-\infty, u_1, u_2, \ldots, u_{n+1}, v_1, \ldots, v_{n+1}, \infty)$ in ascending order, obtaining $\hat{y_0}, \ldots, \hat{y_{2n+3}}$
11: Output $\cup_i [\hat{y_i}, \hat{y_{i+1}}]$, such that $N(\hat{y_i} > r$, where $N(y_i) = \#S_j : [\hat{y_i}, \hat{y_{i+1}}] \subseteq S_j$, where $i = 0, \ldots, 2n$, and $j = 1, \ldots, n+1$.

When $b_i = b_{n+1}$, then $u_i = v_i = \frac{-(a_i + a_{n+1})}{2b_i}$ (from the algorithm), and hence, there is nothing to do. However, our experiments showed that such a scenario was never encountered with the dataset and problem under consideration. When $b_i \neq b_{n+1}$, then:

$$u_i = min(\frac{a_i - a_{n+1}}{b_{n+1} - b_i}, \frac{-(a_i + a_{n+1})}{b_{n+1} + b_i})$$

and

$$v_i = max(\frac{a_i - a_{n+1}}{b_{n+1} - b_i}, \frac{-(a_i + a_{n+1})}{b_{n+1} + b_i})$$

Hence, for the two quantities, u_i and v_i, to be equal (or at least close to each other), we need:

$$\frac{a_i - a_{n+1}}{b_{n+1} - b_i} \approx \frac{-(a_i + a_{n+1})}{b_{n+1} + b_i}$$

where $A = [a_i]$ and $B = [b_i]$ for $i = 1, 2, \ldots, n$ are defined in the algorithm. On

simplification, this implies that we need:

$$2a_{n+1}b_i \approx 2a_i b_{n+1} \Rightarrow \frac{a_{n+1}}{a_i} \approx \frac{b_{n+1}}{b_i}$$

If the above condition is satisfied in a better manner, the efficiency of the CP algorithm for regression will increase, while maintaining the validity.

To verify the above discussion, we studied the values for the above expression in context of the results that were presented in the previous section (Table 4.2). In all of the experiments, it was found that the mean value for $\frac{a_{n+1}}{a_i}$ was 0 across all the test points studied, and the mean value for $\frac{b_{n+1}}{b_i}$ was ∞. Needless to say, since these values are very far apart, the efficiency of the obtained results was very poor. This observation validates our thinking, and corroborates the need to reduce the gap between these ratios.

Achieving the equivalence between the ratios: In order to make the ratios $\frac{a_{n+1}}{a_i}$ and $\frac{b_{n+1}}{b_i}$ equal, let us take a closer look at these quantities. a_{n+1} and b_{n+1} are values from the A and B vectors that pertain to the test data instance under consideration. a_i and b_i, however, are values that pertain to the training data. Since the objective of this work is to build conformal prediction models with maximal efficiency, it would only be possible to learn a distance metric with the training data. In other words, it is easier to obtain better values for a_i and b_i that may achieve the equivalence in the ratios, than be able to obtain better values for a_{n+1} or b_{n+1}. Additionally, when the ridge regression conformal predictors model is trained transductively by adding a new test data point to the dataset, we found that the values of a_i and b_i do not significantly change. In short, a possible solution to achieve the equivalence between the ratios is by optimizing the values of a_i and b_i in the training phase.

The values of a_is and b_is are derived from the following expressions in the

algorithm:

$$C = I - X(X'X + \alpha I)^{-1}X'$$

$$A = C(y_1, y_2, \ldots, y_n, 0)' = (a_1, a_2, \ldots, a_{n+1})$$

$$B = C(0, 0, \ldots, 0, 1)' = (b_1, b_2, \ldots, b_{n+1})$$

Assuming that the data instances in X are normalized (between $[0, 1]$), the construction of the above expressions indicate that the value for a_{n+1} generally hovers close to zero, and the value for b_{n+1} generally hovers close to one. At the same time, the values for both a_i and b_i can be noted to be close to zero. This again explains why the the mean value of the ratio $\frac{a_{n+1}}{a_i}$ is close to zero in our experiments, and the mean value for $\frac{b_{n+1}}{b_i}$ is ∞. Hence, for the ratios to be closer to each other in value, the value of a_is should also become closer to zero, and the value of b_is should become closer to 1.

Now, note that the non-conformity measures for regression [66] are given by the residuals $\Delta = |\mathbf{Y} - \mathbf{X}\mathbf{w}| = |A + B\hat{y}|$. This implies that the a_i values are, in turn, the non-conformity measures or the residuals of the training data instances. In other words, maximizing efficiency in the CP framework for ridge regression, which is equivalent to obtaining lower values of the non-conformity measures a_is, is in fact equivalent to decreasing the error residuals (or the Mean Absolute Error) of the regression function, i.e., a ridge regression function that has lower Mean Absolute Error will maximize efficiency. Hence, in this work, we propose a methodology for learning a distance metric that can provide better regression performance (by minimizing Mean Absolute Error), thereby increasing efficiency of the CP framework.

Metric Learning for Maximizing Efficiency

Metric Learning: A Brief Review

Many of the algorithms in machine learning rely extensively on the definition of a suitable distance metric that is used for computations between data instances. It is known that the distance metric that is selected can play a critical role in the performance of such algorithms. In recent years, studies have shown that a distance metric that is learnt from the data instances directly improves the performance of these algorithms. Such methods of learning distance metrics from data are collectively termed as 'metric learning' techniques.

Given a distance metric denoted by the matrix $A \in \mathbb{R}^{n \times n}$, the generalized definition of the distance between two data points x and y is given by:

$$d_A^2(x,y) = \|x-y\|_A^2 = (x-y)^T A(x-y) \tag{4.1}$$

Yang presented a comprehensive survey of techniques that are used to learn the distance metric matrix A in [3], categorizing these techniques as shown in Figure 4.1. Supervised metric learning methods are used when the labels of training data are available (such as the work of Xing et al. [159]). Unsupervised methods (such as dimensionality reduction techniques [160]) are used otherwise. Within supervised metric learning methods, global approaches ([159] for example) attempt to satisfy a suitable criterion simultaneously for all pairs of data points, while local approaches ([161] for example) are formulated to satisfy such criteria in local neighborhoods. For more details on metric learning techniques, please refer to Yang's work [3]. A Matlab toolbox for distance metric learning is also available in the public domain[1].

[1] http://www.cs.cmu.edu/ liuy/distlearn.htm

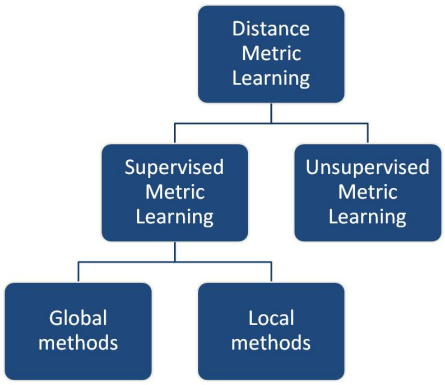


Figure 4.1: Categorization of distance metric learning techniques as presented in [3]

Metric Learning and Manifold Learning: The Connection

In [162], Yang pointed out that from Equation 4.1, one can write:

$$d(x,y) = (x-y)^T A^{\frac{1}{2}} A^{\frac{1}{2}} (x-y) = (A^{\frac{1}{2}}x - A^{\frac{1}{2}}y)^T (A^{\frac{1}{2}}x - A^{\frac{1}{2}}y)$$

$$= (Px - Py)^T (Px - Py)$$

where $P = A^{\frac{1}{2}}$. Hence, learning the metric A is in fact equivalent to learning a linear projection mapping P. It can be thus stated that all linear dimensionality reduction techniques are equivalent to redefining the distance metric between the data points. Considering that many non-linear dimensionality reduction techniques are non-linear versions of their linear equivalents (for example, Locality Preserving Projections [163] is understood to be a linear approximation of Laplacian Eigenmaps [164]), some non-linear dimensionality reduction techniques can also be considered as approximations of metric learning techniques.

In light of the above discussion, several manifold learning techniques, which form a sub-class of dimensionality reduction methods, are equivalent to learning respective distance metrics. Since manifold learning methods have shown promise in recent years when applied to the problem of head pose estimation, we adopt a

manifold learning-based methodology to learn a distance metric that can maximize efficiency.

4.3 Efficiency Maximization in Head Pose Estimation through Manifold Learning

An Introduction to Manifold Learning

The computation of low-dimensional representations of high-dimensional observations like images is a problem that is common across various fields of science and engineering. Techniques like Principal Component Analysis (PCA) are categorized as linear dimensionality reduction techniques, and are often applied to obtain the low-dimensional representation. Other dimensionality reduction techniques like Multi-Dimensional Scaling (MDS) use the dissimilarities (generally Euclidean distances) between data points in the high-dimensional space to capture the relationships between them. In recent years, a new group of non-linear approaches to dimensionality reduction have emerged, which assume that data points are embedded on a low-dimensional manifold in the ambient high-dimensional space. These have been grouped under the term *'manifold learning'*, and some of the most often used manifold learning techniques in the last few years include: Isomap [165], Locally Linear Embedding (LLE) [166], Laplacian Eigenmaps [164], Locality Preserving Projections (LPP) [163], Neighborhood Preserving Embedding (NPE) [167]. A few of these techniques are briefly described below. For more details, please refer to [160] for a review of dimensionality reduction techniques.

Isomap

To capture the global geometry of the data points, Tenenbaum et al. [165] proposed Isomap to compute an isometric low-dimensional embedding of a given set of high-dimensional data points. In this method, the neighbors of a point on the

manifold M are determined, and the neighborhood of each point is represented as a weighted graph G, with each edge characterized by the distance $dx(i,j)$ between the pair of neighboring points, x_i and x_j. The geodesic distances between all pairs of points on the manifold M are estimated by computing their shortest path distance in the graph G. This is done using the Floyds or Djkstraas algorithm, i.e. $dM(x_i,x_j) = min_k \left\{dM(x_i,x_j), dM(x_i,x_k) + dM(x_k,x_j)\right\}$. Classical MDS is then applied to the geodesic distance matrix, deriving an embedding of the data in a low-dimensional Euclidean space that best preserves the estimated intrinsic geometry of the manifold.

Locally Linear Embedding (LLE)

Roweis and Saul [166] proposed the LLE algorithm that embodied the think globally, fit locally paradigm. In this technique, the neighbors of a point of the manifold are determined as for Isomap. The data point is shifted to the origin along with its neighborhood to form a local data matrix Z, and the local covariance $C = Z'Z$ is computed. The linear system $CW = 1$ is solved for the weights W in the neighborhood, which are subsequently normalized. The bottom eigenvectors of a sparse matrix M, constructed as $M = (I-W)'(I-W)$, are used to project the input vectors into the low-dimensional embedding space.

Laplacian Eigenmaps

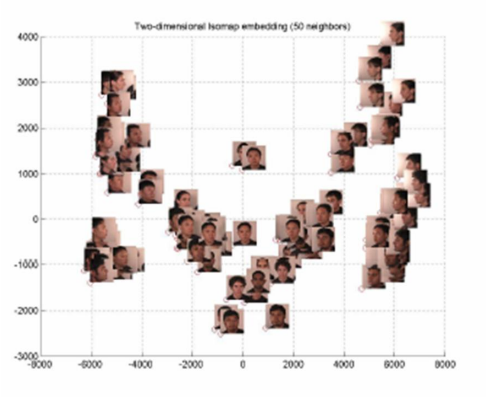
Belkin and Niyogi [164] proposed another geometrically motivated algorithm based on the Laplace-Beltrami operator on a manifold. In this approach, the Laplacian of the graph of the neighborhood of every data point in the feature space is viewed as an approximation to the Laplace-Beltrami operator. A weighted graph is constructed with weight values W drawn from the heat kernel or with a simplistic version, where a weight of unit value is assigned if the nodes are neighbors. The

generalized eigenvector problem $Ly = \lambda Dy$, is solved for the embedding y, where D is the diagonal weight matrix i.e. $D_{ii} = \sum_j W ji$, and $L = D - W$ is the Laplacian matrix.

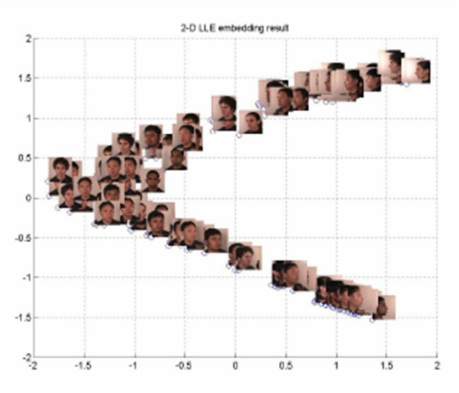
In this work, different poses of the head, although captured in high-dimensional image feature spaces, are visualized as data points on a low-dimensional manifold embedded in the high-dimensional space [136] [138]. The dimensionality of the manifold is said to be equal to the number of degrees of freedom in the movement during data capture. For example, images of the human face with different angles of pose rotation (yaw, tilt and roll) can intrinsically be conceptualized as a 3D manifold embedded in image feature space. We consider face images with pose angle views ranging from $-90\,^{\circ}$ to $+90\,^{\circ}$ from the FacePix database, with only yaw variations. Figure 4.2 shows the 2-dimensional embeddings of face images with varying pose angles from FacePix database obtained with three different manifold learning techniques - Isomap, Locally Linear Embedding (LLE), and Laplacian Eigenmaps. On close observation, one can notice that the face images are ordered by the pose angle. In all of the embeddings, the frontal view appears in the center of the trajectory, while views from the right and left profiles flank the frontal view, ordered by increasing pose angles. This ability to arrange face images by pose angle (which is the only changing parameter) during the process of dimensionality reduction explains the reason for the increased interest in applying manifold learning techniques to the problem of head pose estimation.

Manifold Learning for Head Pose Estimation: Related Work

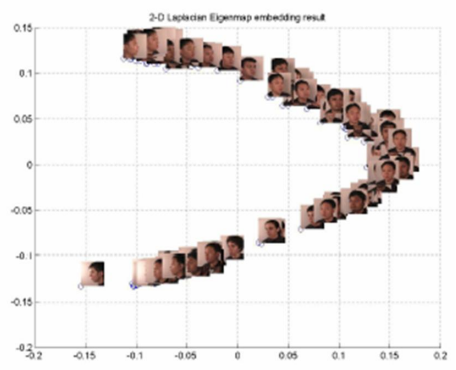
A classification of different approaches to head pose estimation was presented in Table 4.1 in the earlier section. In this section, we discuss approaches to pose estimation using manifold learning that are related to the proposed framework, and



(a) Embedding with the Isomap algorithm



(b) Embedding with the LLE algorithm



(c) Embedding with the Laplacian Eigenmap algorithm

Figure 4.2: Embedding of face images with varying poses onto 2 dimensions

review their performance and limitations.

Since the advent of manifold learning techniques less than a decade ago, a reasonable amount of work has been done using manifold-based dimensionality reduction techniques for head pose estimation. Chen et al. [156] considered multi-view face images as lying on a manifold in high-dimensional feature space. They compared the effectiveness of Kernel Discriminant Analysis against Support Vector Machines in learning the manifold gradient direction in the high-dimensional feature space. The images in this work were synthesized from a 3D scan. Also, the application was restricted to a binary classifier with a small range of head pose angles between $-10\,^{\circ}$ and $+10\,^{\circ}$.

Raytchev et al. [136] studied the effectiveness of Isomap for head pose estimation against other view representation approaches like the Linear Subspace model and Locality Preserving Projections (LPP). While their experiments showed that Isomap performed better than the other two approaches, the face images used in their experiments were sampled at pose angle increments of $15\,^{\circ}$. In the discussion, the authors indicate that this dataset is insufficient to provide for experiments with accurate pose estimation. The least pose angle estimation error in all their experiments was $10.7\,^{\circ}$, which is rather high.

Hu et al. [158] developed a unified embedding approach for person-independent pose estimation from image sequences, where the embedding obtained from Isomap for a single individual was parametrically modeled as an ellipse. The ellipses for different individuals were subsequently normalized through scale, translation and rotation based transformations to obtain a unified embedding. A Radial Basis Function interpolation system was then used to obtain the head pose angle. The authors obtained good results with the datasets, but their approach relied on temporal continuity and local linearity of the face images, and hence was intended for image/video

sequences.

In more recent work, Fu and Huang [135] presented an appearance-based strategy for head pose estimation using a supervised form of Graph Embedding, which internally used the idea of Locally Linear Embedding (LLE). They obtained a linearization of manifold learning techniques to treat out-of-sample data points. They assumed a supervised approach to local neighborhood-based embedding and obtained low pose estimation errors; however, their perspective of supervised learning differs from how it is addressed in this work.

Biased Manifold Embedding for Efficiency Maximization

In this work, a new framework for supervised manifold learning called Biased Manifold Embedding is proposed to address the problem of person-independent head pose estimation [138] [139]. The low regression error obtained through this approach makes it suitable to increase the efficiency of the CP framework, as explained in earlier sections. Before presenting the formulation of this framework, we discuss related efforts that have addressed the problem of supervised manifold learning.

Supervised Manifold Learning: A Review

In the last few years, there have been efforts to formulate supervised approaches to manifold learning. However, none of these approaches have explicitly been used for head pose estimation. In this section, we review the main ideas behind their formulations, and discuss the major novelties in our work, when compared to the existing approaches.

Ridder et al. [168] came up with one of the earliest supervised frameworks for manifold learning. Their framework was centered around the idea of defining a new distance metric for Locally Linear Embedding, which increased inter-class dis-

tances and decreased intra-class distances. This modified distance metric was used to compute the dissimilarity matrix, before computing the adjacency graph which is used in the dimensionality reduction process. Vlassis et al. [169] formulated a supervised approach that was intended towards identifying the intrinsic dimensionality of given data using statistical methods, and using the computed dimensionality for further analysis.

Li and Guo [170] proposed a supervised Isomap algorithm, where a separate geodesic distance matrix is constructed for the training data from each class. Subsequently, these class-specific geodesic distance matrices are merged into a discriminative global distance matrix, which is used for the Multi-Dimensionality Scaling step. Vlachos et al. [171] proposed the WeightedIso method, where the Euclidean distance between data samples is scaled with a constant factor $\lambda (< 1)$, if the class labels of the samples are the same. Geng et al. [172] extended the work from Vlachos et al towards visualization applications, and proposed the S-Isomap (supervised Isomap), where the dissimilarity between two points is defined differently from the regular geodesic distance. The dissimilarity is defined in terms of an exponential factor of the Euclidean distance, such that the intra-class distance never exceeds 1, and the inter-class distance never falls below $1 - \alpha$, where α is a parameter that can be tuned based on the application.

Zhao et al. [173] proposed a supervised LLE (SLLE) algorithm in the space of face images preprocessed using Independent Component Analysis. Their SLLE algorithm constructs these neighborhood graphs with a strict constraint imposed: only those points in the same cluster as the point under consideration can be its neighbors. In other words, the primary focus of the proposed SLLE is restricted to reveal and preserve the neighborhood in a cluster scope.

The approaches to supervised manifold learning discussed above primarily con-

sider the problem from a classification/clustering perspective. In our work, we view the class labels (pose labels) as possessing a distance metric by themselves i.e. we approach the problem from a regression perspective. However, we also illustrate how it can be applied to classification problems. In addition, we show how the proposed framework unifies the existing approaches. The mathematical formulation of the proposed framework is discussed in the next section.

Biased Manifold Embedding: The Mathematical Formulation

In this section, we discuss the mathematical formulation of the Biased Manifold Embedding approach as applied in the head pose estimation problem.

Manifold learning methods, as illustrated in earlier sections, align face images with varying poses by an ordering of the pose angle in the low-dimensional embeddings. However, the choice of image feature vectors, presence of image noise and the introduction of the face images of different individuals in the training data can distort the geometry of the manifold. To ensure the alignment, we propose the Biased Manifold Embedding framework, so that face images whose pose angles are closer to each other are maintained nearer to each other in the low-dimensional embedding, and images with farther pose angles are placed farther, irrespective of the identity of the individual. In the proposed framework, the distances between data points in the high-dimensional feature space are biased with distances between the pose angles of corresponding images (and hence, the name). Since a distance metric can easily be defined on the pose angle values, the problem of finding 'closeness of pose angles is straight-forward.

We would like to modify the dissimilarity/distance matrix between the set of all training data points with a factor of the pose angle dissimilarities between the points. We define the modified biased distance between a pair of data points to be

of the fundamental form:

$$\tilde{D}(i,j) = \lambda_1 \times D(i,j) + \lambda_2 \times f(P(i,j)) \times g(D(i,j)) \tag{4.2}$$

where $D(i,j)$ is the Euclidean distance between two data points x_i and x_j, $\tilde{D}(i,j)$ is the modified biased distance, $P(i,j)$ is the pose distance between x_i and x_j, f is any function of the pose distance, g is any function of the original distance between the data samples, and λ_1 and λ_2 are constants. While we defined this formulation after empirical evaluations of several formulations for the dissimilarity matrix, we found that this formulation, in fact, unifies other existing supervised approaches to manifold learning that modify the dissimilarity matrix.

In general, the function f could be picked from the family of reciprocal functions ($f \in \mathscr{F}_R$) based on an application. In this work, we set $\lambda_1 = 0$ and $\lambda_2 = 1$ in Equation 4.2, function g as the constant function (= 1), and the function f as:

$$f(P(i,j)) = \frac{1}{max_{m,n}P(m,n) - P(i,j)}$$

This function could be replaced by an inverse exponential or quadratic function of the pose distance, for example. To ensure that the biased distance values are well-separated for different pose distances, we multiply this quantity by a function of the pose distance:

$$\tilde{D}(i,j) = \frac{\alpha(P(i,j))}{max_{m,n}P(m,n) - P(i,j)} * D(i,j)$$

where the function α is directly proportional to the pose distance, $P(i,j)$, and is defined in our work as:

$$\alpha(P(i,j)) = \beta * |P(i,j)|$$

where β is a constant of proportionality, and allows parametric variation for performance tuning. In our current work, we used the pose distance as the one-dimensional distance i.e. $P(i,j) = \left|P_i - P_j\right|$, where P_k is the pose angle of x_k.

In summary, the biased distance between a pair of points can be given by:

$$\tilde{D}(i,j) = \begin{cases} \frac{\alpha(P(i,j))}{max_{m,n}P(m,n)-P(i,j)} * D(i,j) & P(i,j) \neq 0, \\ 0 & P(i,j) = 0. \end{cases} \tag{4.3}$$

This biased distance matrix is used for techniques such as Isomap, Locally Linear Embedding (LLE), Locality Preserving Projections (LPP), Neighborhood Preserving Embedding (NPE) and Laplacian Eigenmaps to obtain a pose-ordered low-dimensional embedding. In case of Isomap, the geodesic distances are computed using this biased distance matrix. The LPP, NPE, LLE and Laplacian Eigenmaps algorithms are modified to use these distance values to determine the neighborhood of each data point. Since the proposed approach does not alter the algorithms in any way other than the computation of the biased dissimilarity matrix, it can easily be extended to other manifold-based dimensionality reduction techniques which rely on the dissimilarity matrix.

In Equation 4.3 of the proposed framework, the function $P(i,j)$ is defined in a straightforward manner for regression problems. Further, the same framework can also be extended to classification problems, where there is an inherent ordering in the class labels. An example of an application with such a problem is head pose classification. Sample class labels could be 'looking to the right','looking straight ahead','looking to the left','looking to the far left', and so on. The ordering in these class labels can be used to define a distance metric. For example, if the class labels are indexed by an ordering $k = 1,2,...,n$ (where n is the number of class labels), a simple expression for $P(i,j)$ is:

$$P(i,j) = \gamma \times dist(|i-j|)$$

where i and j are the indices of the corresponding class labels of the training data

samples. The *dist* function could just be the identity function, or could be modified depending on the application.

4.4 Experiments and Results

Experimental Setup

The setup of the experiments conducted in the subsequent sections is described here. All of these experiments were performed with a set of 2184 face images, consisting of 24 individuals with pose angles varying from $-90\,^{\circ}$ to $+90\,^{\circ}$ in increments of $2\,^{\circ}$. The images were subsampled to 32×32 resolution, and two different feature spaces of the images were considered for the experiments. The results presented here include the grayscale pixel intensity feature space and the Laplacian of Gaussian (LoG) transformed image feature space (see Figure 4.3). The LoG transform, which captures the edge map of the face images, was used since pose variations in face images can be considered a result of geometric transformation, and texture information can be considered redundant. The images were subsequently rasterized and normalized.

(a) Grayscale image

(b) Laplacian of Gaussian (LoG) transformed image

Figure 4.3: Image feature spaces used for the experiments

Unlike linear dimensionality reduction methods like Principal Component Analysis, manifold learning techniques lack a well-defined approach to handle out-of-sample extension data points. Different methods have been proposed [174] [175] to

capture the mapping from the high-dimensional feature space to the low-dimensional embedding. We adopted the Generalized Regression Neural Network (GRNN) with Radial Basis Functions to learn the non-linear mapping. GRNNs are known to be a one-pass learning system, and are known to work well with sparsely sampled data. This approach has been adopted by earlier researcher [173]. The parameters involved in training the network are minimal (only the spread of the Radial Basis Function), thereby facilitating better evaluation of the proposed framework. Once the low-dimensional embedding was obtained, ridge regression (regularized least squares regression) [66] was used to obtain the pose angle of the test image. To ensure generalization of the framework, 8-fold cross-validation was used in these experiments. In this validation model, 1911 face images (91 images each of 21 individuals) were used for the training phase in each fold, while all the remaining images were used in the testing phase. The parameters i.e. the number of neighbors used and the dimensionality of embedding were chosen empirically.

Three different sets of experiments were carried out in this work to validate our hypotheses:

- Firstly, the applicability of manifold learning-based techniques over traditional dimensionality reduction techniques such as Principal Component Analysis (PCA) was studied in context of the head pose estimation problem.
- In the second set of experiments, the performance of the proposed Biased Manifold Embedding framework over manifold learning techniques was studied with respect to head pose estimation.
- Lastly, the Biased Manifold Embedding was applied in association with the CP framework for ridge regression, and the measures of efficiency were studied to validate the proposed idea.

Using Manifold Learning over Principal Component Analysis

Traditional approaches to pose estimation that rely on dimensionality reduction use traditional linear techniques such as PCA. However, with the assumption that face images with varying poses lie on a manifold, non-linear dimensionality reduction would be expected to perform better. We performed experiments to compare the performance of manifold learning techniques with Principal Component Analysis. The results of head pose estimation comparing PCA against manifold learning techniques with the experimentation setup described in the previous sub-section are tabulated in Tables 4.3 and 4.4.

Dimension of embedding	Error in pose estimation			
	PCA	Isomap	LLE	Laplacian Eigenmap
10	11.37 °	12.61 °	6.60 °	7.72 °
20	9.90 °	11.35 °	6.04 °	6.32 °
40	9.39 °	10.98 °	4.91 °	5.08 °
50	8.76 °	10.86 °	4.37 °	4.57 °
75	7.83 °	10.67 °	3.86 °	4.17 °
100	7.27 °	10.41 °	3.27 °	3.93 °

Table 4.3: Results of head pose estimation using Principal Component Analysis and manifold learning techniques for dimensionality reduction, in the grayscale pixel feature space

As the results illustrate, while Isomap and PCA perform very similarly, both the local approaches i.e. Locally Linear Embedding and Laplacian Eigenmaps show $3-4$ ° improvement in pose angle estimation over PCA, consistently.

Using Biased Manifold Embedding for Person-independent Pose Estimation

While manifold learning techniques demonstrate reasonably good results for pose estimation over linear dimensionality reduction techniques, we hypothesize that the supervised approach to manifold learning performs better for accurate results with

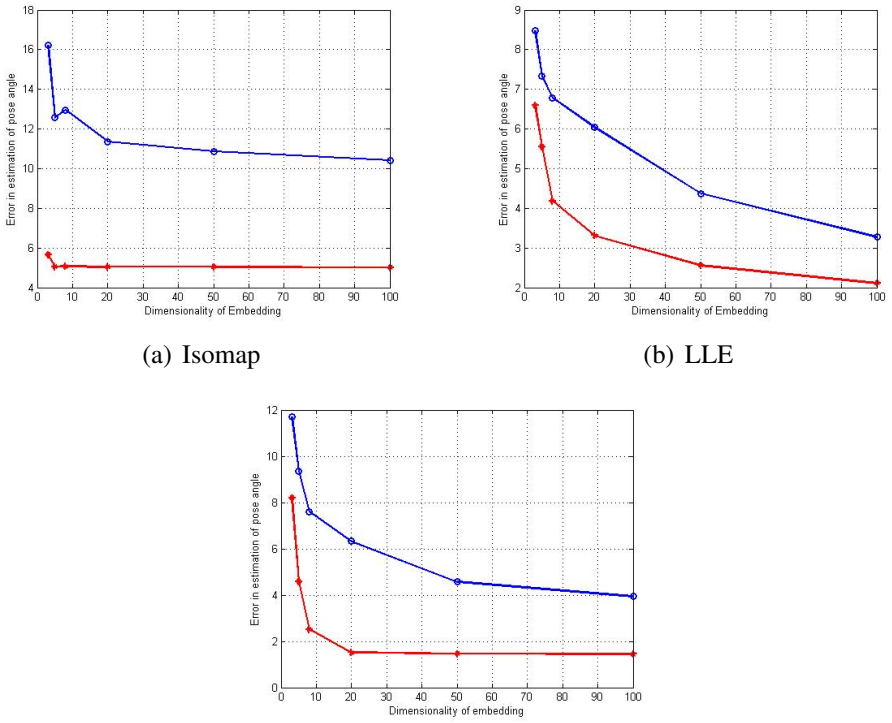
Dimension of embedding	Error in pose estimation			
	PCA	Isomap	LLE	Laplacian Eigenmap
10	9.80 °	9.79 °	7.41 °	7.10 °
20	8.86 °	9.21 °	6.71 °	6.94 °
40	8.54 °	8.94 °	5.80 °	5.91 °
50	8.03 °	8.76 °	5.23 °	5.23 °
75	7.92 °	8.47 °	4.83 °	4.89 °
100	7.78 °	8.23 °	4.31 °	4.52 °

Table 4.4: Results of head pose estimation using Principal Component Analysis and manifold learning techniques for dimensionality reduction, in the LoG feature space

person-independent pose estimation. In our next set of experiments, we evaluate this hypothesis. The error in the pose angle estimation process is used as the criterion for the evaluation, since this can indirectly affect the efficiency when applied in association with the CP framework.

The proposed BME framework was applied to face images from the FacePix database, and the performance was compared against the performance of regular manifold learning techniques. These experiments were performed against global (Isomap) and local (Locally Linear Embedding and Laplacian Eigenmaps) approaches to manifold learning. The error in the estimated pose angle (against the ground truth from the FacePix database) was used to evaluate the performance.

The results of these experiments are presented in Figures 4.4 and 4.5. The blue line indicates the performance of the manifold learning techniques, while the red line stands for the performance from the Biased Manifold Embedding approach. As evident, the error significantly drops with the proposed approach. All of the approaches perform better with the LoG feature space, as compared to using plain grayscale pixel intensities. This corroborates the intuitive assumption that the head pose estimation problem relies more on the geometry of face images, and the tex-



(a) Isomap

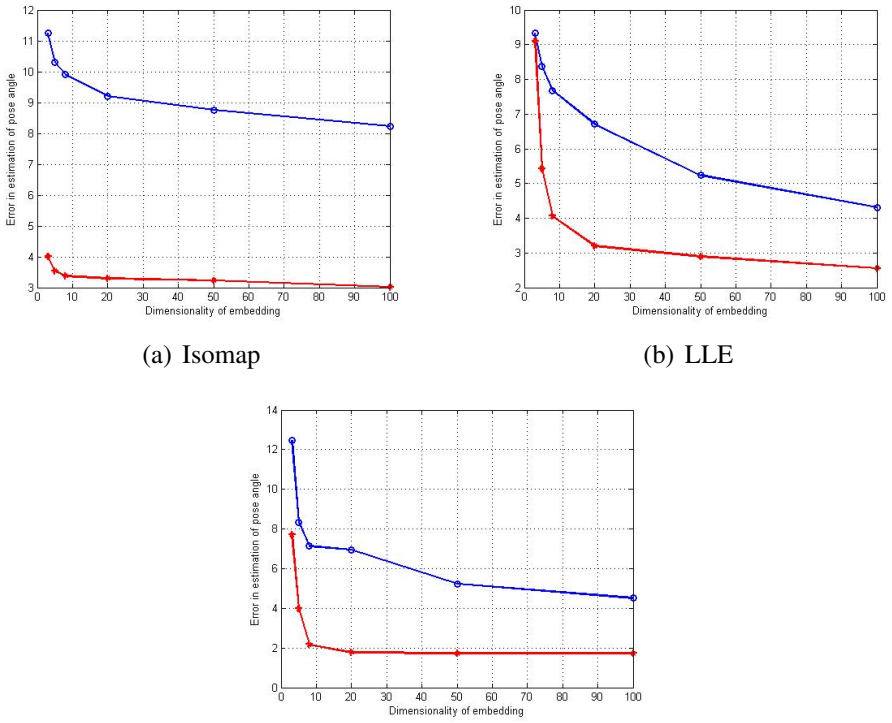
(b) LLE

(c) Laplacian Eigenmap

Figure 4.4: Pose estimation results of the BME framework against the traditional manifold learning technique with the grayscale pixel feature space. The red line indicates the results with the BME framework

ture of the images can be considered redundant. However, we believe that it would be worthwhile to perform a more exhaustive analysis with other feature spaces as part of our future work. Also, it is clear from the error values obtained that the BME framework substantially improves the head pose estimation performance, when compared to other manifold learning techniques or Principal Component Analysis.

It can also be observed that the results obtained from the local approaches i.e. Locally Linear Embedding and Laplacian Eigenmaps far outperform the global approach, viz. Isomap. Considering that Isomap is known to falter when there is topological instability [4], the relatively low performance with both the feature spaces suggests that the manifold of face images constructed from the FacePix database



(a) Isomap

(b) LLE

(c) Laplacian Eigenmap

Figure 4.5: Pose estimation results of the BME framework against the traditional manifold learning technique with the Laplacian of Gaussian (LoG) feature space. The red line indicates the results with the BME framework

may be topologically unstable. In reality, this would mean that there are face images which short-circuit the manifold in a way that the computation of geodesic distances is affected (See Figure 4.6). An outlier could short-circuit the geometry of the manifold, and destroy its geometrical structure. In such a case, global approaches like Isomap fail to find an appropriate low-dimensional embedding. There have been recent approaches to overcome the topological instability by removing critical outliers in a pre-processing step [175].

Comparison with other related pose estimation work: In comparing related approaches to pose estimation which have different experimental design criteria, the results are summarized below in Table 4.5. The results obtained from the BME

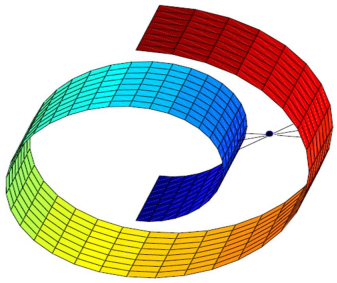


Figure 4.6: Example of topological instabilities that affect Isomap's performance (Illustration taken from [4])

framework match the best results so far obtained by [135], considering face images with pose angle intervals of 1 °. The best results are obtained when BME is used with Laplacian Eigenmap. When LLE or Isomap is used, the error goes marginally higher and hovers about 3 °.

Method	Pose Angle Estimation Error/Accuracy	Notes
Fisher Manifold Learning [156]	About 3 °	Face images only in $[-10\,^\circ, 10\,^\circ]$ interval
Kernel PCA + Support Vector Machines [152]	97%	Face images only in 10 ° intervals. This was framed as a classification problem of identifying the pose angle as one of these intervals
Isomap [136]	About 11 °	Face images sampled at 15 ° increments
LPP [136]	About 15 °	Face images sampled at 15 ° increments
LEA [135]	About 2 °	Best results so far
Proposed BME using Laplacian Eigenmap	About 2 °	Results similar to [135]
Proposed BME using Isomap, LLE	About 3 °	-

Table 4.5: Summary of head pose estimation results from related approaches in recent years

Having shown the superior regression performance of the Biased Manifold Embedding framework in terms of the Mean Absolute Error, our final set of experiments studied the performance of this method for improving efficiency, when applied with the CP framework. In these experiments, Biased Manifold Embedding was applied on the data points to obtain a low-dimensional embedding that has a more conducive ordering of face images for pose estimation. As stated earlier, this step can be perceived as imposing a new distance metric on the data. These low-dimensional embeddings were then used as input data for the CP framework algorithm in the regression context (Algorithm 4, Section 4.2).

Our experiments in the earlier subsection indicated that LLE and Laplacian Eigenmaps (LE) performed better than Isomap in terms of the mean pose estimation error. Hence, in this study, these two methods were studied in the broader context of the CP framework, and compared with the performance of applying ridge regression conformal predictors directly. In addition, the linear approximations of LLE and LE - Neighborhood Preserving Embedding (NPE) and Locality Preserving Projections (LPP) respectively - were also used for the study. An extended dataset containing the face images of 30 users (instead of 24 users in the previous experiments) was used in these experiments. Further, the Laplacian of Gaussian feature space was used in this study, since it demonstrated better performance over the plain grayscale pixel intensities. Other experimental conditions remained the same as described in Section 4.4.

The results of these experiments are presented in Table 4.6 and Figure 4.7. In this table, the baseline method refers to applying the ridge regression CP algorithm without any dimensionality reduction. Note that the most desirable result is a lower

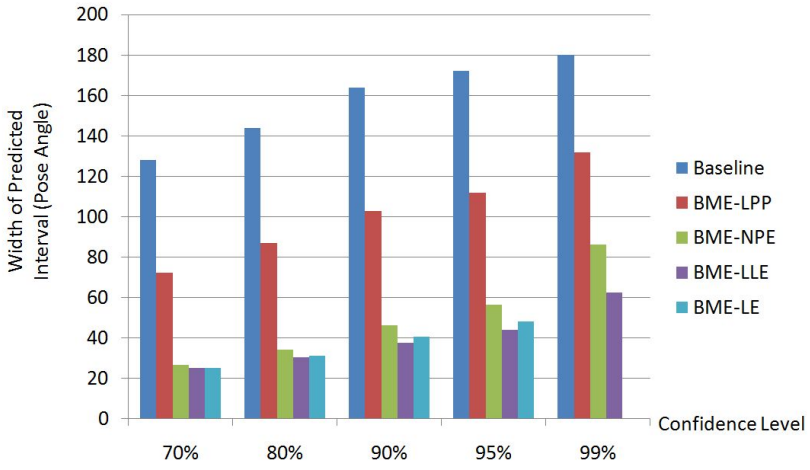


Figure 4.7: Summary of results showing the width of the predicted interval using the proposed Biased Manifold Embedding (BME) framework in association with 4 manifold learning techniques: LPP, NPE, LLE and LE

value of the mean width of the predicted interval, while maintaining calibration in the percentage of errors at each confidence level. As the table illustrates, the mean width of the prediction region has significantly reduced when applying Biased Manifold Embedding, while maintaining calibration at each confidence level. In some cases, the percentage of reduction in the mean width is as high as 300%. Although there are minor statistical fluctuations observed in the percentage of errors, this is not an issue of concern, and can be expected in real-world data. These results validate our hypothesis, corroborating the usefulness of the proposed method for improving efficiency.

We further studied the relationship between the observed results and the values of the ratios discussed earlier in the chapter (which motivated this approach), and these observations have been noted in Table 4.7. It is evident that when compared with the baseline, the mean absolute values for the ratios $\frac{a_{n+1}}{a_i}$ and $\frac{b_{n+1}}{b_i}$ have decreased significantly, once again validating our conceptual framework. However, these values indicate that there is substantial room for improvement, and this will

Method	Percentage of Errors at Confidence Level				
	70%	80%	90%	95%	99%
Baseline	30.04	20.07	9.96	5.02	1.03
LLE	31.65	20.7	9.89	4.29	1.32
LE	29.93	20.29	9.45	4.51	1.14
NPE	30.88	20.7	10.37	5.68	2.16
LPP	31.14	21.06	11.36	6.37	1.76

Method	Mean Width of Prediction Region at Confidence Level				
	70%	80%	90%	95%	99%
Baseline	128	144	164	172	180
LLE	25.1	30.46	37.83	44.2	62.53
LE	25.25	31.42	40.63	48.1	70.13
NPE	26.74	34.24	46.23	56.25	86.07
LPP	72.36	86.83	102.69	111.71	131.92

Table 4.6: Results of experiments studying efficiency when Biased Manifold Embedding is applied along with the CP framework for head pose estimation. Note that baseline stands for no dimensionality reduction applied, LLE: Locally Linear Embedding, LE: Laplacian Eigenmaps, NPE: Neighborhood Preserving Embedding, LPP: Locality Preserving Projections

be the focus of our future work.

Method	Mean Absolute $\frac{a_{n+1}}{a_i}$ Value	Mean Absolute $\frac{b_{n+1}}{b_i}$ Value
Baseline	0	∞
LLE	5.02	32116.69
LE	6.115	110972.3
NPE	0.641	32707.78
LPP	0.206	18919.24

Table 4.7: Values of the ratios for a_is and b_is in the CP ridge regression algorithm for each of the methods studied

4.5 Discussion

Biased Manifold Embedding: A Unified View of Other Supervised Approaches

The proposed Biased Manifold Embedding framework can be considered as a unified representation of other existing supervised manifold learning approaches. In the next few paragraphs, we discuss briefly how the existing supervised approaches to manifold learning are special cases of the Biased Manifold Embedding framework. Although this discussion is not directly relevant to the pose estimation problem, this shows the broader appeal of this idea.

Ridder et al. [168] proposed a supervised LLE approach, where the distances between the samples are artificially increased if the samples belonged to different classes. If the samples are from the same class, the distances are left unchanged. The modified distances are given by:

$$\Delta' = \Delta + \alpha \times max(\Delta)\Lambda, \alpha \in [0,1]$$

Going back to Equation 4.2, we arrive at Ridder et al's formulation by choosing $\lambda_1 = 1$, $\lambda_2 = \alpha \times max(\Delta)$, function $g(D(i,j)) = 1 \forall i,j$, and function $f(P(i,j)) = \Lambda$.

Li and Guo [170] proposed the SE-Isomap (Supervised Isomap with Explicit Mapping), where the geodesic distance matrix is constructed differently for intra-class samples, and is retained as is for inter-class data samples. The final distance matrix, called the discriminative global distance matrix G, is of the form:

$$G = \begin{bmatrix} \rho_1 G_{11} & G_{12} \\ G_{21} & \rho_2 G_{22} \end{bmatrix}$$

Clearly, this representation very closely resembles the choice of parameters we have chosen in our pose estimation work. In Equation 4.2, the formulation of Li and Guo would simply mean choosing $\lambda_1 = 0$, $\lambda_2 = 1$, function $f(P(i,j)) = 1$, and function

$g(D(i,j))$ can be defined as:

$$g(D(i,j)) = \begin{cases} D(i,j) & P(i) \neq P(j), \\ \rho_i \times D(i,j) & P(i) = P(j). \end{cases}$$

The work of Vlachos et al. [171] - the WeightedIso method - is exactly the same in principle as the work of Li and Guo. For data samples belonging to the same class, the distance is scaled by a factor $\frac{1}{\alpha}$, where $\alpha > 1$; else, the distance is left undisturbed. This can be exactly formulated as discussed above for Li and Guo. The work of Geng et al. [172] is based on the WeightedIso method, and the authors extended the WeightedIso method with a different dissimilarity matrix (which would just mean a different definition for $D(i,j)$ in the proposed BME framework), and parameters to control the distance values.

Zhao et al. [173] formulated the S-LLE (supervised LLE) method, where the distance between points that belonged to different classes was set to infinity i.e. the neighbors of a particular data point had to belong to the same class as the point. Again, this would be rather straight-forward in the BME framework, where the function $g(D(i,j))$ can be defined as:

$$g(D(i,j)) = \begin{cases} \infty & P(i) \neq P(j), \\ D(i,j) & P(i) = P(j). \end{cases}$$

The proposed BME framework can, hence, be considered as providing a unified view of existing supervised manifold learning approaches.

Finding Intrinsic Dimensionality of Face Images

An important component of manifold learning applications is the computation of the intrinsic dimensionality of the dataset provided. Similar to how linear dimensionality reduction techniques like PCA use the measure of captured variance to

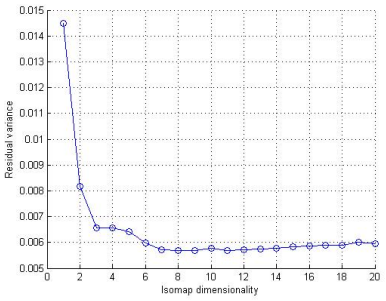
arrive at the number of dimensions, manifold learning techniques are dependent on knowing the intrinsic dimensionality of the manifold embedded in the high-dimensional feature space.

We performed a preliminary analysis of the dataset to extract its intrinsic dimensionality, similar to what was performed in [165]. Isomap was used to perform non-linear dimensionality reduction on a set of face images from 5 individuals. Different pose intervals of the face images were selected to vary the density of the data used for embedding. The residual variances after computation of the embedding are plotted in Figure 4.8. The sub-figures illustrate that most of the residual variance is captured in one dimension of the embedding. This goes to prove that there is only one dominant dimension in the dataset. As the pose intervals used for the embedding becomes lesser i.e. the density of the data becomes higher, this observation is even more clearly noted. The data captured in the FacePix database have pose variations only along one degree of freedom (the pitch), and this result corroborates the fact that these face images could be visualized as lying on a low-dimensional (ideally, one-dimensional) manifold in the feature space.

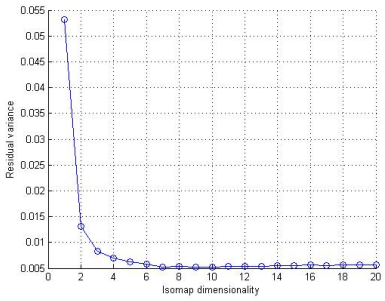
Experimentation with Sparsely Sampled Data

Manifold learning techniques have been known to perform poorly on sparsely sampled datasets [160]. Hence, in our next set of experiments, we propose that the BME framework, through supervised manifold learning, performs reasonably well even on sparse samples, and evaluate this hypothesis.

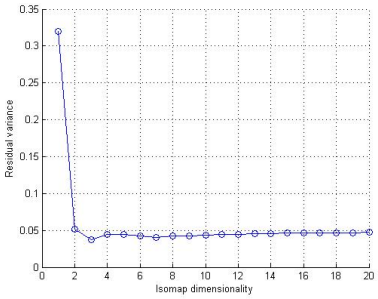
In these experiments, we sampled the available set of face images sparsely (by pose angle) and used this sparse sample of the face images dataset for training, before testing with the entire dataset. In these experiments, face images of all the 30 individuals in the FacePix database were used. The set of training images included



(a) Face images with 5 ° pose angle intervals



(b) Face images with 2 ° pose angle intervals



(c) Face images with 1 ° pose angle intervals

Figure 4.8: Plots of the residual variances computed after embedding face images of 5 individuals using Isomap

face images in pose angle intervals of 10 ° i.e. only 19 out of the total 181 images for each individual were used in the training phase. Subsequently, the number of training images (total number of images is 5430) was progressively reduced in steps to observe the performance. These experiments were carried out for Isomap, LLE and Laplacian Eigenmaps (LE) for both the feature spaces. To maintain uniformity of results and to aid comparison, all these trials embedded the face images onto a 8-dimensional space, and 50 neighbors were used for constructing the embedding (as in the earlier section). The results are presented in Tables 4.8 and 4.9. Note the results obtained with BME and without BME for Isomap and Laplacian Eigenmap in both these tables. The results show significant reduction in error. However, the

results for LLE do not reflect this observation.

Number of training images	Error using Isomap		Error using LLE		Error using LE	
	w/o BME	w/ BME	w/o BME	w/ BME	w/o BME	w/ BME
570	12.13 °	3.26 °	5.95 °	5.88 °	10.27 °	3.84 °
475	11.70 °	6.01 °	6.58 °	6.95 °	9.47 °	3.71 °
380	8.19 °	7.61 °	6.47 °	6.72 °	9.59 °	4.72 °
285	8.39 °	8.75 °	6.36 °	6.71 °	9.12 °	5.61 °
190	8.75 °	8.58 °	6.77 °	7.03 °	10.05 °	7.76 °
95	11.27 °	9.22 °	9.43 °	8.45 °	15.44 °	14.54 °

Table 4.8: Results from experiments performed with sparsely sampled training dataset for each of the manifold learning techniques with (w/) and without (w/o) the BME framework on the grayscale pixel feature space. The error in the head pose angle estimation is noted

Number of training images	Error using Isomap		Error using LLE		Error using LE	
	w/o BME	w/ BME	w/o BME	w/ BME	w/o BME	w/ BME
570	10.63 °	3.19 °	8.76 °	7.99 °	9.01 °	3.57 °
475	12.08 °	3.73 °	8.08 °	7.63 °	8.56 °	3.99 °
380	11.34 °	6.40 °	8.16 °	8.48 °	8.47 °	5.00 °
285	13.96 °	6.66 °	8.14 °	8.49 °	9.30 °	6.69 °
190	15.46 °	6.96 °	8.72 °	8.68 °	12.27 °	8.84 °
95	11.93 °	8.59 °	8.77 °	8.77 °	30.17 °	15.79 °

Table 4.9: Results from experiments performed with sparsely sampled training dataset with (w/) and without (w/o) the BME framework on the LoG feature space

The results validate our hypothesis that the BME framework performs better even with sparsely sampled datasets. With Isomap and Laplacian Eigenmap, the application of the BME framework improves the performance of pose estimation substantially. However, we note that Locally Linear Embedding performed as well

even without the Biased Manifold Embedding framework. This suggests that in tasks of unsupervised learning (like clustering), where there are no class labels to supervise the learning process, Locally Linear Embedding may be a good technique to apply for sparsely sampled datasets.

Limitations of Manifold Learning Techniques

Over the last few years, the increased application of manifold learning techniques has also resulted in identification of some limitations of these methods [160] [176]. While all these techniques capture the geometry of the data points in the high-dimensional space, the disadvantage of this family of techniques is the lack of a projection matrix to embed out-of-sample data points after the training phase. This makes the method more suited for data visualization, rather than classification/regression problems. However, the advantage of these techniques to capture the relative geometry of data points enthuses researchers to adopt this methodology to solve problems like head pose estimation, where the data is known to possess geometric relationships in a high-dimensional space.

These techniques are known to depend on a dense sampling of the data in the high-dimensional space. Also, Ge et al. [177] noted that these techniques do not remove correlation in high-dimensional spaces from their low-dimensional representations. The few applications of these techniques to pose estimation have not exposed the limitations yet - however, from a statistical perspective, these generic limitations intrinsically emphasize the requirement for the training data to be distributed densely across the surface of the manifold. In real-world applications like pose estimation, it is highly possible that the training data images may not meet this requirement. This brings forth the need to develop techniques that can work well with training data on sparsely sampled manifolds too.

4.6 Summary

In this chapter, we proposed an approach for improving efficiency in the Conformal Predictions framework for regression by indirectly learning a distance metric through supervised manifold learning. A novel framework called the Biased Manifold Embedding was proposed as a result for to person-independent head pose estimation. Under the credible assumption that face images with varying pose angles lie on a low-dimensional manifold, non-linear dimensionality reduction based on manifold learning techniques possesses strong potential for face analysis in biometric applications. We compared the proposed framework with regularly used approaches like Principal Component Analysis and other manifold learning techniques, and found the results to be reasonably good for head pose estimation. Importantly, the proposed supervised manifold learning approach provided encouraging results for improving the efficiency of ridge regression conformal predictors.

In future work, we wish to study the usefulness of formulating a methodology to explicitly learn a distance metric to maximize efficiency. In addition, we plan to implement the inductive version of the CP framework (which is known to have substantially lesser demands on computational overhead) as part of a wearable platform to perform real-time pose classification from a live video stream, to study its applicability in real-world scenarios. Further, as manifold learning techniques continue to be applied in pose estimation and similar applications, it becomes imperative to carry out an exhaustive study to identify the kind of image feature spaces that are most amenable to manifold-based assumptions and analysis.

4.7 Related Contributions

Similar to the previous chapter, there were other contributions that were made to address related problems in the development of a Social Interaction Assistant, which

laid the context for this work on head pose estimation. This assistive device is intended to help individuals with visual impairments to experience enriched interactions in daily life. These contributions included a systematic requirements analysis for this device [81], as well as the design and development of a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind [178] [87]. In addition, conceptual approaches to human-centered multimedia computing using inspirations from disabilities and deficits have also been suggested [82] [179] [180].

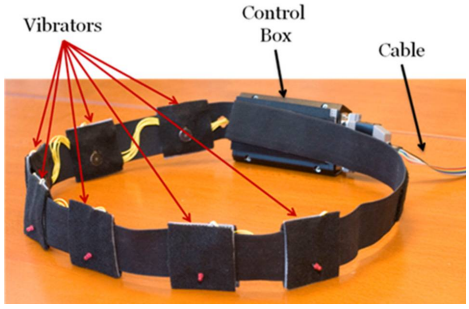


Figure 4.9: A first prototype of the haptic belt for the Social Interaction Assistant

Chapter 5

CONFORMAL PREDICTIONS FOR INFORMATION FUSION

Sources of multimedia data have grown rapidly in the last few decades, resulting in the generation of data from different modalities, sensing technologies and processing techniques. The relevance of information fusion methods has increased over the years, and these methods have elicited keen interest from researchers in recent years. However, with different data sources and modeling methods, there is an additional factor that determines the success of information fusion methods - handling the uncertainties of each of these sources and models, and being able to associate a level of confidence to the final fused result. This chapter addresses this aspect of information fusion using the reliable measures of confidence from the Conformal Predictions (CP) framework.

Handling different uncertainty frameworks to derive a single reliable belief measure is a challenging task. Several theories have been proposed and studied in this regard, and some of these theories are listed in Section 5.1. Since the Conformal Predictions framework provides reliable measures of confidence under the calibration property and can be generalized to a wide variety of classification and regression methods, extending the framework to information fusion contexts can potentially result in valuable and impactful contributions. The development of a methodology to extend the CP framework to information fusion (for both classification and regression) and the validation of the calibration property under these settings is the objective of the contribution in this chapter. This methodology is validated in a classification setting on the multimodal person recognition problem, and in regression on the saliency prediction problem.

5.1 Background and Motivation

In estimating the uncertainty when there are multiple modalities and classifiers involved, there have been several theories to aggregate the evidence developed over the years. Some of these approaches are listed below:

- Dempster-Shafer theory [181]
- Bayesian theory [182]
- Possibility theory [183]
- Fuzzy integrals [184]
- MYCIN uncertainty factors [185]
- DSmT combination [186]
- Belief functions theory [187]
- GESTALT system [188]

In a more specific survey on reliability in information fusion, Rogova and Nimier [189] reviewed existing approaches that have attempted to incorporate the reliability of sources in the results of information fusion methods. In their detailed account of methods that handle reliability co-efficients in decision fusion, they classified uncertainty frameworks commonly employed as combinatorial functions in fusion systems into:

- *Bayesian methods*, which include probabilistic methods that use the prior probability, likelihood and posterior probabilities in the system.
- *Evidential methods*, such as the Dempster-Shafer theory of evidence [181], and the transferable belief model [187].

- *Possibility and fuzzy methods*, where most of the combination rules are based on t-norms and t-conorms, the fuzzy translation of intersection and union.

While the methods listed above have become popular over the years, the choice of approach in a given application domain is generally empirical or even heuristic. Further, it is not possible to establish desired properties of a confidence measure such as validity/calibration (or in some cases, generalizability to existing classification and regression methods), and there have been extensive criticisms of approaches such as the Dempster-Shafer theory in literature [190]. One aspect of the limitations of existing theories is the lack of a well-defined 'calibration' property when multiple sources are involved, as in information fusion settings. In this work, we specifically focus on addressing this issue in information fusion problems for classification and regression contexts.

The fusion of information from multiple sources can happen at different levels using different methods, as summarized in Figure 6.3 [5]. Dasarathy [191] categorized these approaches as *data-level fusion* (where data is combined), *feature-level fusion* (where features are extracted from the data in different modalities separately, and these features are then combined), and *decision-level fusion* (where the fusion happens at the decision-making level). Data-level fusion and feature-level fusion are sometimes addressed together as *early fusion*, whereas decision-level fusion is also called *late fusion*. Similar to DeCampos et al. [192], we are interested in this work to be able to associate a confidence measure to every single test prediction uniquely. In most early fusion techniques, the weights of the individual components are pre-set or learnt from training data; and hence, such weights remain fixed for new samples. In order to ensure that we associate a confidence measure uniquely to every test sample, we approach the problem from a late fusion perspec-

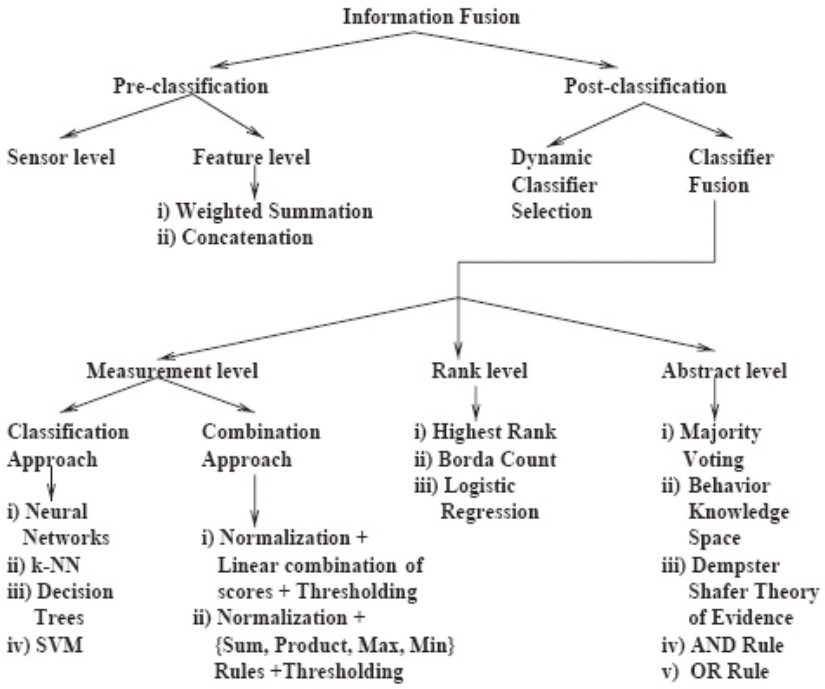


Figure 5.1: An overview of approaches to fusion, with details of methods in classifier fusion, also called decision-level fusion [5]

tive. Decision-level fusion (or late fusion) is the focus of this work, and a discussion of how early fusion can be addressed within the CP framework is presented in the next section.

Rationale and Significance: Confidence Estimation in Information Fusion

Extending the CP framework to information fusion can be approached from two distinct perspectives: *pre-fusion* and *post-fusion*. These terms are used in a sense similar to how information fusion methods are categorized into data-level, feature-level and decision-level fusion [191]. These terms are explained below:

- *Pre-fusion:* Measures of confidence are computed with respect to each classifier (or regressor), and these measures of confidence are then combined in a second stage to give a combined confidence value.

- *Post-fusion:* The output scores of the involved classifiers (or regressors) are combined using standard techniques before a single measure of confidence is computed for the ensemble.

The *post-fusion* computation of confidence is relatively straightforward, when compared to the *pre-fusion* case. Vovk et al. [38] suggested that a suitable non-conformity measure can be defined after the outputs of each of the classifiers have been combined. For example, in the case of an ensemble classifier such as boosting, the non-conformity measure can be defined as:

$$\sum_{t=1}^{T} \alpha_t B_t(x,y)$$

where B_t are the weak classifiers, and α_t are the weights learnt by the boosting algorithm. This measure can be directly plugged into the CP framework to obtain calibrated measures of confidence with the desired properties. The issues discussed in the previous section with respect to maximizing efficiency remain as challenges, but there is no explicit challenge related to the combination of classifiers.

However, the *pre-fusion* approach, which is the focus of this work, has not been addressed earlier, and can be of value in application contexts. This approach has more challenges, since a measure of confidence is computed for every classifier (for convenience, this discussion is presented for classification, although equally relevant to regression methods) in the system, and an appropriate methodology to combine these measures of confidence needs to be identified. Fundamental questions such as: 'should the confidence measures be combined into a single value of confidence', 'what would calibration mean when measures of confidence are combined', and 'given a final single measure of confidence, would it be possible to elicit any information about the individual confidence measures in each of the classifiers' arise in this context, and need to be addressed.

Significance of a pre-fusion approach: To understand why the pre-fusion case is required when the post-fusion case can be solved far more easily, let us consider the problem of multimodal person recognition, i.e., the task of recognizing the identity of an individual using, say, both face and speech data. In the post-fusion case, the computed confidence value would be applicable only to the combined outputs from the face and speech classifiers respectively. However, if the user would like to understand which of the modalities resulted in errors (so that appropriate corrections can be made, possibly in the form of an additional training phase for that modality), it would be essential to have a measure of confidence for each of the modalities, and understand how they contributed to the net confidence. This example illustrates the need for a pre-fusion perspective in combining confidence values in multiple classifier systems.

In other words, confidence can be viewed as being computed at an *entity level* and at an *attribute level* in an information fusion context, where an entity such as a person is understood to be made up of several attributes such as face and speech. While the post-fusion approach computes only the entity-level confidence, the pre-fusion approach can provide an attribute-level confidence and an entity-level confidence, thus providing a higher value to the end user.

We now outline our approach to combine conformal predictors from multiple classifiers (or regression methods) for information fusion.

5.2 Methodology: Conformal Predictors for Information Fusion

Before presenting our methodology for this work, we review key challenges that generally need to be addressed in the design of multiple classifier Systems, and motivate the methodology adopted in this work.

Key Challenges

Ranawana and Palade [193] presented a comprehensive survey of the challenges that need to be addressed in the design of multiple classifier systems. Although these challenges have been reviewed from the perspective of classification methods, they are pertinent to regression contexts too. These challenges can be broadly categorized as:

- Selection of appropriate classifiers
- Selection of a suitable combinatorial function
- Selection of a representative topology for classifier integration

Selection of Appropriate Classifiers

While considering several classifiers to solve a problem, it is important that each of these classifiers are selected with a purpose. There have been several studies in earlier work on the desirable properties of such classifiers. For example, Lam [194] identified complementarity, orthogonality and independence as essential traits in the selection of classifiers. Applying the CP framework to such systems does not raise any additional challenges, since the non-conformity measure can be appropriately defined for each of the selected classifiers. The rest of the procedures remain the same.

Selection of a Suitable Combinatorial Function

A large number of existing efforts that address challenges in information fusion have focused on this specific issue, i.e., how do we combine the outputs of the classifiers? There have been a variety of approaches to achieve this task, and a summary of combinatorial functions in existing literature is presented below:

- Linear combination methods (like SUM and PROD),

- Non-linear combination methods (like majority voting),
- Statistical methods (like Dempster-Shafer theory, or Bayesian functions), and
- Computationally intelligent methods (like neural networks or genetic algorithms).

The CP framework has the desirable property to provide calibrated measures of confidence when applied in association with a single classifier or a single classifier ensemble (such as boosting). However, it cannot be guaranteed if the CP framework will provide calibrated outputs when combinatorial functions are applied to individual classifiers. The study and identification of combinatorial functions that can maintain calibration when the p-values from individual conformal predictors are combined is the objective of this work.

Selection of Topologies for Classifier Integration

Ranawana-Palade [193] and Lam [194] identified four kinds of topologies in which multiple classifiers can be combined:

- Conditional topology, where one classifier performs the classification and another classifier is selected if the first classifier fails to correctly identify the presented data.
- Hierarchical (Serial) topology, where classifiers are applied in succession one after another.
- Multiple (Parallel) topology, where all classifiers operate in parallel on the input (or parts of it), and the results are then pooled to obtain a consensus result.
- Hybrid topology, where the best classifier for a given input is selected and used.

Most often, classifiers are combined in the parallel topology, and in this case, it may be easy to define an appropriate non-conformity measure for the combined multiple classifier system. However, when classifiers are combined using topologies such as conditional or hybrid, it may not be straightforward to define a non-conformity measure - and even if a non-conformity measure can be defined, it may not be a straightforward task to understand how calibration would be defined, given the different classifiers. This remains a significant challenge in the computation of confidence for information fusion contexts, and will be an important direction of our future work.

We now describe our method to combine p-values of different conformal predictors to provide a second level of fused conformal predictions.

Combining P-values from Multiple Classifiers/Regressors

Given a new test data instance, the Conformal Predictions framework outputs a p-value for every class label in a classification context (or for every relevant interval in regression), as described in Chapter 2. Without any loss in generalization to regression settings, we continue this discussion from a classification perspective for the sake of convenience.

Given a classifier, a new test data instance is tested for each class label as a hypothesis test, thereby resulting in a p-value for each class label. When there are multiple data sources (such as face and speech for person recognition, or different feature spaces from a single image for saliency prediction) and when there is a separate classifier that is used to learn a model for each of these data sources, it is evident that for each class label, we obtain a set of p-values using the CP framework, where each entry corresponds to a single data source. Hence, if the hypothesis tests for the different data sources can be combined into a single combined hypothesis

test, it is possible to generate p-values at the fusion level.

Combining p-values from multiple hypothesis tests is a well-studied problem in statistics, and many methods have been proposed in this regard to obtain a single resultant p-value. Some of the oldest methods that combine p-values from multiple tests are listed below:

- Tippett's method [195]
- Fisher's method [196]
- Wilkinson's method [197]
- Liptak's method [198]
- Lancaster's method [199]
- Edgington's method [200]
- Mudholkar and George [201]
- Weighted combination methods [202] [203]

Over the last decade, there has been a renewed interest among researchers in statistics to study procedures for meta-analysis, especially in the use of combinations of p-values to produce a single overall test of hypotheses. A summary of methods when the individual tests are dependent can be found in [204], and a comparison of methods to combine p-values of independent tests was carried out by Loughin [205]. In this work, it is assumed that the individual tests are independent. This is a reasonable assumption for applications such as multimodal biometrics where the face and speech data can be considered to be independent to a large extent.

When the individual tests are considered to be independent, the general setup, as stated by Loughin, is as follows. The combined null hypothesis, H_0, is that each

of the individual null hypotheses (say $H_{01}, H_{02}, \ldots, H_{0N}$) is true. The combined alternative, H_A, is that at least one of the alternatives (say $H_{A1}, H_{A2}, \ldots, H_{AN}$) is true. A p-value p_i, $i = 1, \ldots, k$, is given for each of the individual tests. These are combined into a new statistic $C = C(p_1, \ldots, p_k)$, which is used to test the combined hypothesis.

There are two kinds of methods that are generally used to combine p-values: *quantile combination methods* and *order statistic methods* [205], each of which is described below:

- *Quantile combination methods:* In such methods, some parametric Cumulative Distribution Function (CDF), F, is selected, and the p-values, p_is, are transformed into distributional quantiles, $q_i = F^{-1}(p_i)$ where $i = 1, 2, \ldots, k$ for each of the class labels. These q_is are subsequently combined as $C = \sum_i q_i$, and the p-value of the combined test H_0 is computed from the sampling distribution of C. Examples of CDFs used in these methods include chi-square [196] [199], standard normal [198], uniform [200] and logistic [201].
- *Order statistic methods:* These methods use the fact that under the null hypothesis H_0, the p_is can be reordered as $p_{(i)}$s such that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$ represent order statistics from a $U(0,1)$ distribution (Note that a p-value is assumed to be a uniformly distributed random variable on the interval $[0,1]$). Then, a combining function C is defined as $C = p_{(r)}$ for r such that $1 \leq r \leq k$. Common examples of order statistic methods are the minimum p-value (when $r = 1$ [195]) and the maximum p-value (when $r = k$ [205]).

In the study conducted by Loughin [205], the author concluded that the standard normal quantile combination method is probably the best suited for general use.

In yet another method proposed by Jost [6], the combined p-value is computed as follows. For n experiments or analyses, one can create an n-dimensional unit hypercube and plot the point $(p_1, p_2, p_3, \ldots, p_m)$ representing the p-values p_i of each of the m data sources or corresponding classifiers. A surface of points with the same probability as this point can then be established (Figure 5.2). Since the p-values are independent probabilities (under the null hypothesis), the individual probabilities can be multiplied to give the probability of obtaining this set of p-values. The set of points whose probability is equal to that of the given set of p-values is then the hyper-surface:

$$(x_1 \times x_2 \times x_3 \times \cdots \times x_m) = k$$

where $k = (p_1 \times p_2 \times p_3 \cdots \times p_m)$, the product of the given set of p-values. By

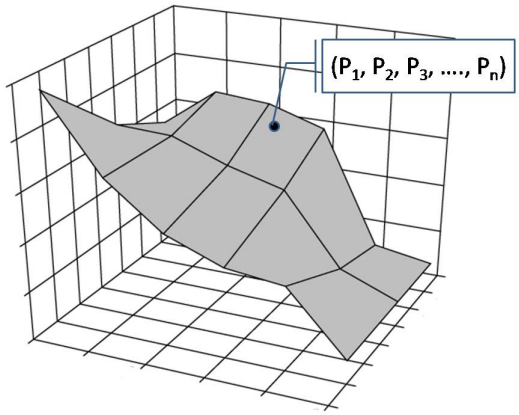


Figure 5.2: A surface of points with the same probability as the point $(p_1, p_2, p_3, \ldots, p_m)$ representing the p-values p_i of each of the m classifiers or data sources (Illustration taken from [6])

definition of a p-value in statistics, we need the probability of getting a set of p-values as extreme or more extreme than the given set. Therefore, we need to find the volume under this hyper-surface. Because p-values are uniformly distributed random variables, and because the total volume of the cube equals 1, the volume

under the surface directly gives the probability of obtaining a set of p-values as extreme or more extreme than the given set. The volume integral depends only on k (the product of the given set of p-values) and m, the number of p-values under consideration. The overall significance level, for the case of two p-values, is then given by:

$$k - k\ln k$$

And for m tests, the combined significance level is given by:

$$k\sum_{i=0}^{m-1}\frac{(-\ln k)^i}{i!} \tag{5.1}$$

This formula to combine the results from each of our features to obtain a final p-value for each class label.

In a completely different approach to addressing this issue, DeCampos et al. [192] recently used a Support Vector Machine (SVM) that takes the p-values of each of the data sources as an input vector, and then uses the CP framework to obtain p-values at the fusion level. They adopted this approach for combining different image features in saliency detection.

Based on inspirations gathered from earlier work, we identified methods of three different kinds to study in this work. These methods have been listed below.

- Quantile Combination Approaches

  - *Standard Normal Function (SNF):* This was found to be the most suitable for general use in an earlier study [205]. In this approach, we compute the inverse of the normal CDF using the p-values obtained from the individual classifiers and thus compute $q_i = F^{-1}(p_i)$ for $i = 1, 2, \ldots, k$. C is then obtained as $\sum_i q_i$, and the normal CDF is again used as the sampling distribution to compute the p-values at the fusion level.

- *Non-conformity Aggregation (NCA):* The non-conformity measure values computed in the CP framework can be viewed as the 'test statistic' leading to the computation of the p-values for each class label. Hence, instead of assuming a quantile function, F^{-1}, and then computing the q_i values, the non-conformity measures themselves can be used as the q_is. Similar to the previous approach, C is then obtained as $\sum_i q_i$, and the combined C values are then used as non-conformity measures at the fusion level to compute the p-values using the standard CP framework procedure (Equation 2.7 in Chapter 2).

- *Extended Chi-Square Function (ECF):-* Fisher proposed the chi-square quantile combination method to combine the p-values of independent tests in [196]. However, Jost [6] stated that when Fisher's derivation for the chi-square statistic is solved further analytically, the result is the expression in Equation 5.1 (described earlier in this section). Hence, we call this the Extended Chi-Square Function (ECSF) approach in our work. The chi-square CDF was also recommended was by Loughin in their study of such methods for general use along with the standard normal function [205].

- Order Statistic Approaches

  - *Minimum Order Statistic (MIN):* The minimum of the p-values for each classifier, the 1st order statistic $p_{(1)}$, is used in this method. Although the usefulness of this combination method has been doubted, this method provided the best results among the order statistic methods [205] and hence, is used in our work.

  - *Maximum Order Statistic (MAX):* The maximum of the p-values for each classifier, the k^{th} order statistic $p_{(k)}$, is used in this method. This method

is understood to perform well only when all null hypotheses are equally false, and has not been recommended for general use. However, for the sake of completeness, we include this method in our study.

- Learning Approaches

  - *k-Nearest Neighbor (KNN):* Similar to [192], we provide the p-values computed from the individual data sources as input to a k-NN classifier, and the CP framework is applied to the k-NN to obtain the p-values at the fusion level.

The aforementioned 6 methods (SNF, NCA, ECF, MIN, MAX, KNN) are used in this work to combine the conformal predictors from individual classifiers and regressors. It is also assumed that the p-values from each of the tests have equal importance. However, there are methods that combine p-values in a weighted manner such as in [202] [203], and these methods can be adopted depending on the need of an application context. The methodology proposed to extend the CP framework to information fusion contexts is validated on two different real-world applications: multimodal person recognition (classification setting), and saliency prediction (regression setting). The experiments and results obtained in each of these application domains are described individually below in Sections 5.3 and 5.4 respectively.

5.3 Classification: Multimodal Person Recognition

We begin this section with a discussion of related work in information fusion within the specific context of multimodal biometrics.

Related Work

The field of biometrics has been extensively studied over the last two decades, and several surveys of research efforts have been presented in recent years [206] [207]

[208] [209]. Since this work is based on combining the face and speech modalities, a survey of various approaches that have attempted to use both these modalities for person recognition is presented in Table 5.1. As the table demonstrates, a wide range of fusion techniques have been tried over the years in this particular application. However, none of these methods can guarantee calibrated measures of confidence in the fused results. In this dissertation's work, we intend to study how the CP framework can be extended to address this problem, by considering each of these classifiers (for face and speech) as an independent hypothesis test, and subsequently combining the p-values obtained from each of these hypothesis tests using the methods described in the previous section.

Experiments and Results

The VidTIMIT and Mobio datasets were used in this work for our study, and these datasets have been described earlier in Section 2.2 in Chapter 2. Support Vector Machines (SVM) was used as the classifier of choice for face data in both these datasets. The Lagrange multipliers obtained while training a SVM are a straightforward choice to consider as non-conformity scores, as pointed out by Vovk et al. [38]. The Lagrange multipliers' values are zero for examples outside the margin, and lie between 0 and a positive constant, C, for examples on and within the margin, thereby providing a natural monotonic measure of *non-conformity* w.r.t. the corresponding class.

The classifier for the speech data is based on a Gaussian Mixture Model (GMM) framework. The speech signal was downsampled to 8 KHz and a short-time 256-pt Fourier analysis is performed on a 25ms Hamming window (10ms frame rate). Every frame log-energy was tagged as high, medium and low (low and 80% of the medium log-energy frames were discarded). The magnitude spectrum was trans-

Trait/ Modality	Algorithms Used
Face	- Elastic Bunch Graph Matching (EBGM) algorithm [210] [211]; - Neural Network using feature vector extracted for each eye [212]; - Similarity metric using features extracted from eye, nose and mouth regions [213]; - Grayscale feature with *k*-NN, decision tree and logistic regression [214]; - Grayscale feature using SVM (second order polynomial kernel) [215]; - Local appearance based models [216]; - Dynamic link architecture [217]; - Multiscale morphological operations [218]; - Feature vector based on DCT with FaceIt [219]; - Fisherfaces [220]; - Principal Component Analysis (PCA) [221] [222] [223]
Speech	- Linear Prediction Cepstral Coefficients (LPCC) with HMMs [210] [211]; - Mel Frequency Cepstral Coefficients (MFCC) [216] [221] [219] [222] [220]; - Gaussian Mixture Models of frequency filtering coefficients [217] [223]; - Segmenting speech signal with wavelet convolution [212]; - Vector quantization of acoustic parameter space [213]; - *k*-NN, decision tree and logistic regression [214]
Fusion	- Bayesian approach with SVMs [210] [211]; - Logical AND [212]; - Weighted geometric average [213]; - Linear weighted summation [215] [221] [218] [222]; - Adaptive modality weighting model called Cumulative Ratio of Correct Matches (CRCM) [216]; - Modality weighting based on estimates of the probability density function of scores under Gaussian assumption [217]; - Cascaded approach where the outputs are weighted by the confidence scores [219]; - Weighting modality scores where weight is proportional to recognition rate [223]

Table 5.1: Summary of approaches in existing work towards fusion of face and speech-based person recognition

formed to a vector of Mel-Frequency Cepstral Coefficients (MFCCs). Further, a feature warping algorithm is applied on the obtained features. Afterwards, a gender-dependent 512-mixture GMM Universal Background Model was initialised using k-means algorithm and then trained by estimating the GMM parameters via the Expectation Maximization algorithm. Target-dependent models were then obtained with MAP (maximum a posteriori) speaker adaptation. Finally, the score computation followed a hypothesis test framework. For more implementation details, please refer to [90] [224] [93]. To adapt this to the CP framework, the negative of the likelihood values generated by the GMM were used as the non-conformity scores, as suggested by Vovk et al. in [38].

Calibration of Errors in Individual Modalities

Before studying the performance of our methodology in combining the p-values of the individual classifiers, the calibration of errors when the CP framework is applied to the individual modalities (as outlined above) was observed. These results are shown in Figure 5.3 and Figure 5.4 for the Mobio dataset. A similar result was observed for the VidTIMIT dataset also. These figures validate that the number of errors for both these modalities are calibrated individually at each of the confidence levels. While there are statistical fluctuations seen in these figures, we believe that this was due to the low number of speech samples that were available for study in this work.

Calibration in Multiple Classifier Fusion

Each of the six methods outlined in Section 5.2 were used to combine the p-values obtained from the CP framework in each of the face and speech modalities. The combined p-values were subsequently used to get a new set of predictions, whose 'conformity' is then studied in this experiment. The obtained results are presented

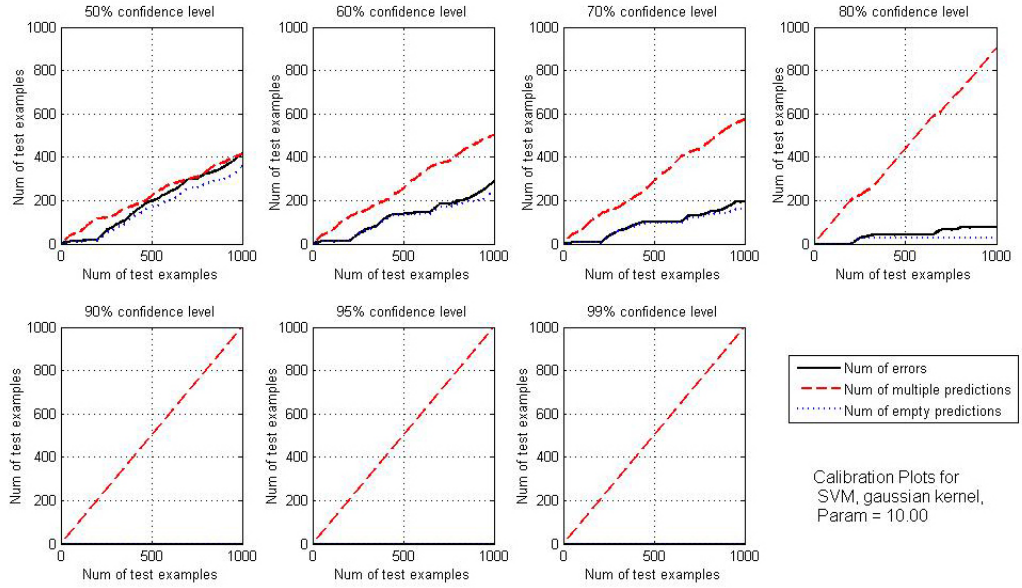


Figure 5.3: Results obtained on face data of the Mobio dataset (SVM classifier)

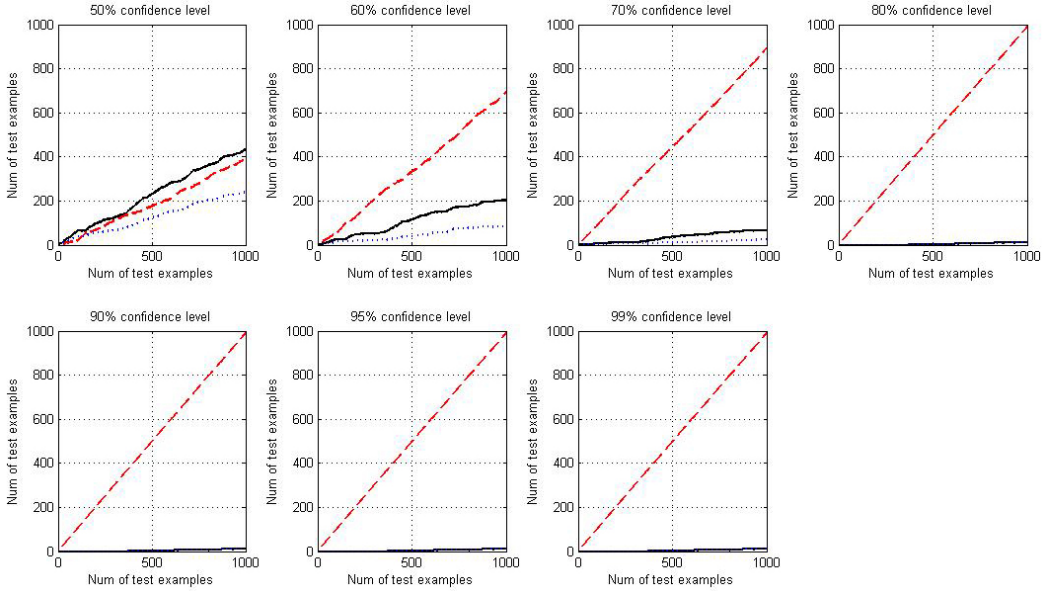


Figure 5.4: Results obtained on speech data of the Mobio dataset (GMM classifier)

Combination Method	Percentage of Errors at Confidence Level						
	50%	60%	70%	80%	90%	95%	99%
SNF	50.9%	40.1%	29.9%	16.1%	1.3%	1%	1%
NCA	55.2%	42.1%	29.7%	5%	1%	1%	1%
ECF	47%	29.8%	9.4%	1%	1%	1%	1%
MIN	76.7%	62.6%	42.4%	4.2%	1%	1%	1%
MAX	20.2%	5.6%	0%	0%	0%	0%	0%
KNN	19.5%	3.6%	1%	1%	1%	1%	1%

Table 5.2: Fusion results on the VidTIMIT dataset. The combination methods have been described in Section 5.2. For k-NN, $k = 5$ provided the best results which are listed here

in Tables 5.2 and 5.3. Evidently, while none of the methods provide ideal results, quantile combination methods (Standard Normal Function, Non-conformity Aggregation and Extended Chi-square Function) provide the highest promise. This is in agreement with the conclusions made in the earlier study conducted by Loughin in [205]. It is possible that these methods may provide better calibration with a more rigorous empirical testing with other ranges of parameter values. However, considering that the number of errors, while not strictly calibrated, does not exceed the specified confidence level, we conclude that these quantile combination methods can be used to combine p-values to extend conformal predictors to information fusion contexts for classification.

5.4 Regression: Saliency Prediction

In Chapter 2 (Section 2.2), we presented the background and objective of this application, i.e. to predict the saliency of every pixel in a radiological image (X-ray image) by learning a model of regions of interest from human eye movements. As in the previous section, we begin with a discussion of prior work in this regard.

Combination Method	Percentage of Errors at Confidence Level						
	50%	60%	70%	80%	90%	95%	99%
SNF	51.7%	42.2%	32.4%	23.6%	11.5%	2.7%	1%
NCA	51.2%	42.8%	21.1%	4.9%	1%	1%	1%
ECF	49%	36.6%	24.7%	10%	1%	1%	1%
MIN	76%	61.5%	48.9%	28%	1%	1%	1%
MAX	16.7%	4.5%	0.2%	0%	0%	0%	0%
KNN	1%	1%	1%	1%	1%	1%	1%

Table 5.3: Fusion results on the Mobio dataset. The combination methods have been described in Section 5.2. We obtained the same results for different values of k in k-NN

Related Work

In earlier work, saliency detection has been studied from two different perspectives: visual attention modeling and interest point detection. Figure 5.5 provides a high-level illustrative summary of the different kinds of approaches that have been used in this context. Visual attention in humans is driven by bottom-up, as well as top-down approaches. Bottom-up attention is driven by regions having distinct features, and is independent of the task, or the context of the scenario. On the other hand, top-down saliency is specific to a context where the user is trying to locate something in particular. Approaches to detect interest points make use of predetermined filters that measure saliency based on specific artifacts such as motion, and corners. The following section reviews the related work from each of these perspectives, and also discusses prior work in the use of human eye movements.

Visual Attention Modeling Methods

Bottom-up Saliency: Research in visual attention modeling has primarily focused on bottom-up saliency. Bottom-up attention is driven by regions having salient stimuli. Such approaches involve algorithms that detect regions in an image/video

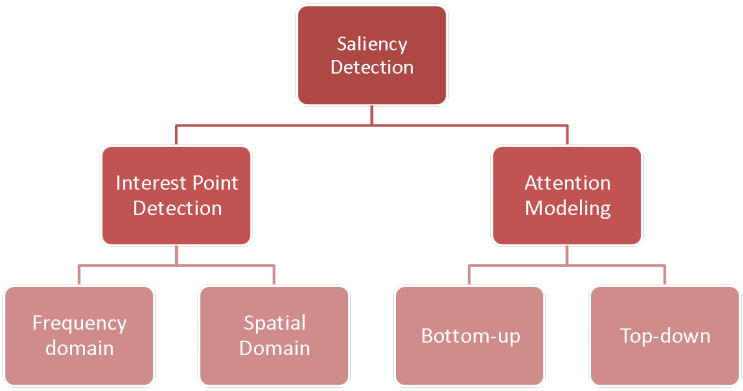


Figure 5.5: Prior work in Saliency detection

that have distinctive features, and is independent of the context of the video. Most computational frameworks that model bottom-up attention are based on the feature integration theory, as explained in [225]. This theory explains the visual search strategy in humans. It proposes that several features are used to obtain individual feature maps that are then integrated to provide the final measure of saliency. The most popular framework to model bottom-up attention was proposed by Itti et al. in [7], as illustrated in Figure 5.6. This model was built on the architecture that was proposed by Koch and Ullman in [226] which is based on the concept of a saliency map that indicates the visual saliency of every pixel in an image. Another approach was proposed by Gao and Vasconcelos in [227]. In their formulation, they equate saliency to discriminability. Although their approach also uses the concept of obtaining different feature maps and combining them into a single saliency map, the filters they use are more suited to locate regions that have discriminative features. Stentiford [228][229] proposed a measure of saliency that depended on the dissimilarity between neighborhoods in an image. This was also linked to the notion of *self-similarity* or a fractals approach. For more details on these approaches, a com-

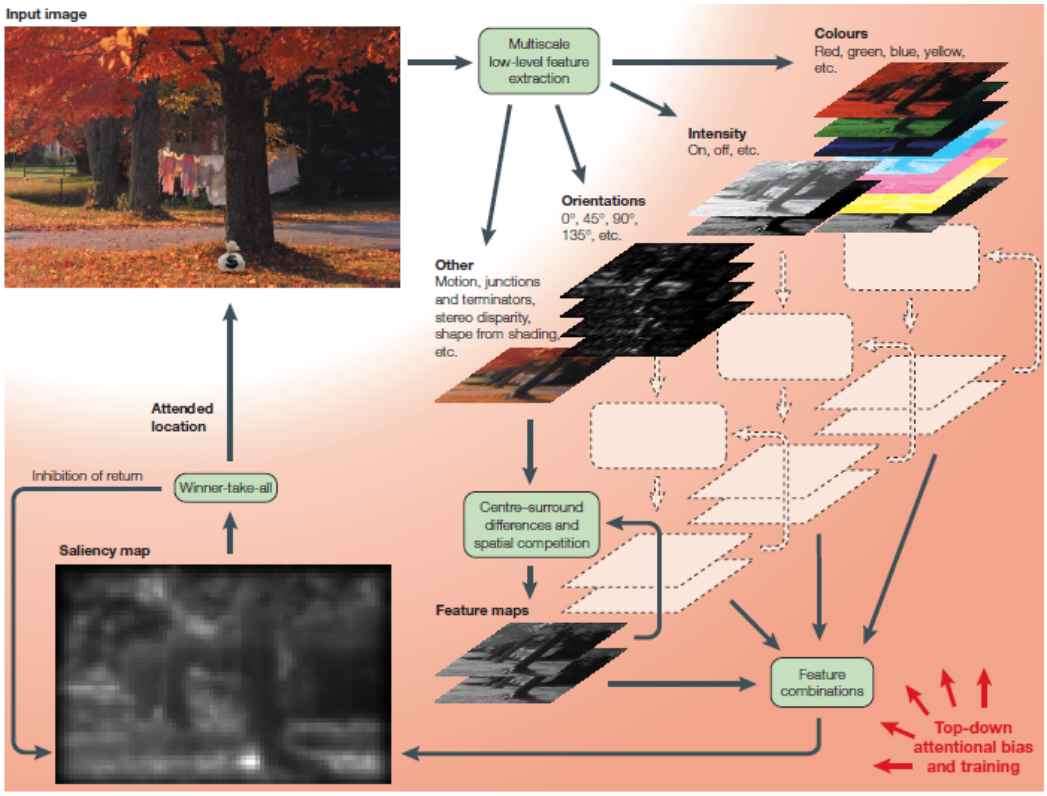


Figure 5.6: Framework used by Itti and Koch in [7] to model bottom-up attention (Illustration taken from [8])

prehensive survey of approaches to model bottom-up attention in humans that aim to extract such conspicuous regions is presented in [230].

Top-down Saliency: Unlike bottom-up saliency, top-down saliency is driven by user intent. There has been limited work done in this regard [231][232][233][234]. Existing approaches that model top-down saliency use a specific goal or task, such as object recognition. Figure 5.7 illustrates such an approach, where two different saliency maps are derived for a single image based on car and person recognition respectively.

Similar to their approach to bottom-up saliency, Gao and Vasconcelos in [231] presented a discriminative saliency based approach for top-down processing. Here,

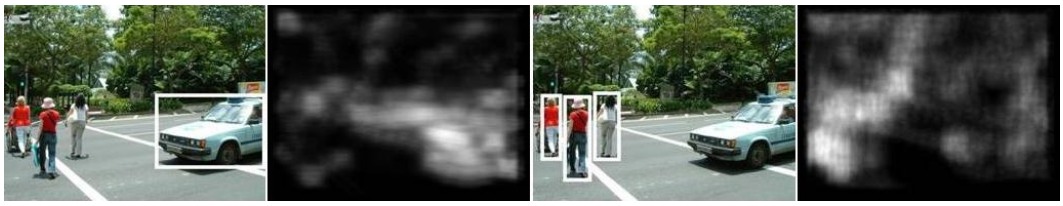


Figure 5.7: Top-down saliency maps derived using recognition based approaches (Illustration taken from [9])

they defined saliency based on features that are most discriminative for a given class of objects. Navalpakkam and Itti, in [232] investigated the use of top-down strategies to select optimal cues in a predetermined search task for objects in a scene, and came up with a model to maximize the saliency of the target with respect to its distractors. In this work, top-down saliency can be viewed as being defined from a different perspective. The task, or the goal, is subtly indicated by the regions fixated by the users while viewing the images, rather than pre-determining user intent in a scene.

Interest Point Detection Methods

Interest point detection refers to identifying a set of pixel locations in images based on a certain saliency, or 'interestingness' measure. The filters or functions used to detect salient locations are chosen so as to respond to certain artifacts such as corners, textures, or motion. Different approaches use different sets of filters to measure the saliency of a pixel. In [235], Lowe proposed the SIFT algorithm to find 'key points' and their corresponding descriptors that are invariant to scale and orientation. Another popular approach to detect interest points is the Harris corner detector [236] that uses a measure based on a second moment matrix to compute the 'cornerness' of a pixel. Kadir and Brady [237] proposed an approach that measures interestingness based on the information content. In their approach, Shannon en-

tropy is used to measure the complexity of a pixel, using the probability distribution of a descriptor extracted from the region around it. This is evaluated across various scales. The final saliency measure of a pixel is calculated based on the scales that exhibit a peak in entropy, as well as a high gradient in the probability distribution.

Another set of pure vision-based approaches are directed towards detecting saliency in images from their Fourier spectrum. These approaches make use of the $\frac{1}{f}$ law [238][239] which describes the statistics of natural images. It states that the amplitude spectrum of the Fourier Transform of an image is inversely proportional to the frequency, i.e. on a log scale, the amplitude spectrum of images is approximately a straight line. Hou and Zhang [240] use this property to define saliency based on the spectral residual of an image. This is calculated as a difference between the log spectrum of an image with its averaged spectrum. The saliency map is constructed using inverse Fourier Transform of this spectral residual. Wang and Li also make use of this property to detect saliency in color images in [241]. However, in their approach they propose a two step approach, where a coarse saliency map is obtained in the first step based on the spectral residual of the image. In the second step, this map is refined and thresholded to obtain a binary saliency map. Another approach in this regard [242] proposes the use of phase spectrum, over the amplitude spectrum of the Fourier Transform, and argue that it provides better results with lesser computations.

In the spatio-temporal domain, there have been two categories of approaches to find interest points. One of them seeks to extend the algorithms that exist in 2-D spatial domain to the temporal domain. Examples of these approaches include the 3-D Harris corner detector [243], 3D SIFT descriptor [244], and the work of Oikonomopoulos et al. [245], which is an extension of the work done by Kadir and Brady in [237] into the third dimension. On the other hand, there have been

approaches that have been specifically designed to identify interestingness in the spatio-temporal domain. A popular algorithm in this regard is the periodic detector, that was proposed by Dollar et al. in [246]. In their approach, a train of pixels in a temporal neighborhood is considered, and their response to a quadrature pair of 1-D Gabor filters in the temporal domain is used to measure saliency. Such an approach associates saliency with any kind of complex motion in a video. Another approach suggested by Ke et al. in [247] makes use of volumetric features and video optical flow to detect motion. This is based on the rectangular features used by Viola and Jones [248].

Interest point detectors differ from attention modeling approaches in that the filters used may not necessarily be biologically driven by the human visual system. These detectors are approached purely from a computer vision perspective. The attention modeling frameworks, on the other hand, are influenced by neurological and cognitive perspectives. Both these approaches, however, do not take the context of the scene into account. Saliency is only a measure of the distinctiveness of the neighborhood of a pixel in terms of texture, motion and other such features. Instead, the approach in this work based on human eye movements computes saliency as a measure of user interest in a given class of images or videos.

Human Eye Movement as Indicators of User Interest

Eye-tracking is the procedure of tracking the position of the eye gaze of a user. One of the earliest uses of eye-tracking was in the field of psychology in understanding how text is read. Researchers analyzed the variations in fixation and saccade durations with line spacing and difficulty of textual content. Eye-tracking was also used to understand scene and art perception. In more recent times, eye-tracking is being increasingly used in commercial and research applications ranging from

Human Computer Interaction (HCI) to medical research to marketing. In recent work on advertising and web applications, eye-trackers are used to monitor the eye-movements of consumers as they view specific websites. This data is statistically analyzed and used to determine the effectiveness of an advertisement, strategize the location of a product, learn the features that catch the attention of an eye, and so on. Duchowski [249] presented an exhaustive survey of various applications that have used eye tracking, with a specific focus on its use in interactive applications.

Eye-tracking has also been used in the field of computer vision and machine learning. Granka et al. in [250] used eye tracking to understand how users react to results provided by an internet search engine, and gain insight into user behavior. Salojarvi et al. [251] investigated the use of eye movements to study relevance in information retrieval. Oyekoya and Stentiford [252][253] conducted experiments in image retrieval that indicated the fact that eye gaze is attracted by regions of interest in images. They found that eye tracking data can be used to retrieve images faster than random selection.

Use of Eye-tracking in Related Work: There has been limited work in detecting salient regions in images using human eye movements. Kienzle et al. [254][255] used human eye movements to learn a model to detect bottom-up saliency in images. They recorded eye-gaze data of users as they were viewing 200 natural images, and built a classifier to learn the image patterns that resulted in high visual saliency. Pixel intensities in rectangular image patches were used as the feature vectors for the classifier. The results indicated that the performance of their model was comparable to other bottom-up approaches. More recently, Judd et al. [256] used eye movements to learn saliency in images. They used a larger dataset of 1003 randomly selected landscape and portrait images to collect eye tracking data of users. As stated in their work, their methodology is very closely related to the

work of Kienzle et al. In their approach, in addition to having low-level features descriptors such as color and contrast, the classifier is also trained on mid-level and high-level features. These include horizon line detectors since most objects are on the ground and humans look for objects, face and person detectors, as well as the distance from the center of the image.

For videos, Kienzle et al. [257] extended their work to detect spatio-temporal interest points in videos. The dataset used for the training comprised of arbitrary videos sampled from a movie. The eye movements of users were recorded as they watched these videos, and a classifier was trained to learn the features corresponding to regions viewed by the users. The feature descriptor for the learning model was based on the periodic detector [246], where a sequence of pixels in the temporal neighborhood of a pixel (pixels having the same spatial coordinates in neighboring frames) is used. However, in their approach, the filter coefficients of the temporal filters were learnt based on the eye movements of users, instead of using a 1-D Gabor filter. More recently, Nataraju et al. [234][258] proposed an integrated approach to combine top-down and bottom-saliency in news videos using human eye movements. In this approach, the top-down saliency was learnt from eye movements of users, and the bottom-up saliency was adapted from the popular Itti's model [7].

In all the aforementioned methods that have been studied for detecting saliency in images or videos, there has been not much work in providing regions of saliency depending on a user-specified confidence level. Very recently, de Campos et al. [192] used the CP framework to detect saliency using image features for image retrieval. However, in this case, this work did not learn the saliency using human eye movements, nor did it study the maintenance of calibration under information fusion. We now present the studies conducted in our work to predict saliency in radiological images using human eye movements, and to extend the CP framework to informa-

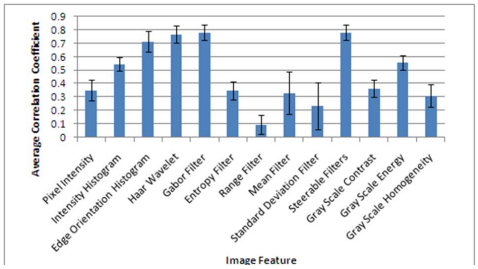
tion fusion in a regression setting.

Experiments and Results

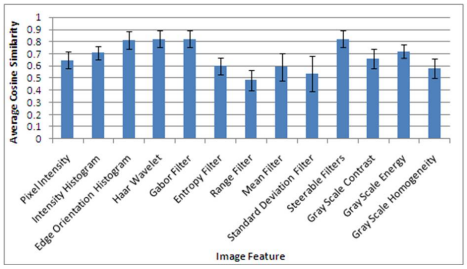
Selecting Image Features for Saliency Prediction

Using the radiological images data described in Chapter 2 (Section 2.2), a study was first performed to learn what types of features catch the eye experienced radiologists when reading chest x-rays for diagnostic purpose [259]. This information can then be used to produce saliency maps that predict what regions of each image might be most interesting to radiologists. The following features were used in this study: Pixel intensity, Intensity histogram, Edge orientation histogram, Haar wavelet, Gabor filter, Entropy filter, Range filter, Mean filter, Standard deviation filter, Steerable filters, Grayscale contrast, Grayscale energy, and Grayscale homogeneity.

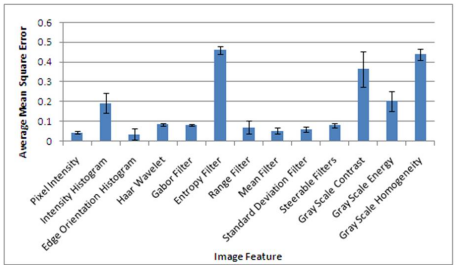
For each of the feature types listed above, Support Vector Regression (SVR) was used to find the single mapping function that was able to produce 20 predicted saliency maps that were (collectively) most similar to the corresponding 20 aggregate empirical saliency maps for the 20 chest x-rays. The SVM-KM Matlab toolbox [79] was used to obtain the SVR model in this work. The similarity between the predicted saliency map and the corresponding aggregate saliency map was computed using three popular similarity metrics: (1) the correlation coefficient, (2) the cosine metric, and (3) the mean-square error metric. Note that since each of these 3 metrics measures the similarity between a pair of n-dimensional vectors, the 2D predicted saliency map and the corresponding 2D aggregate empirical saliency map were each 'unwrapped' to produce a vector. Each of the aforementioned similarity metrics produced a set of 20 similarity/error values - one for each chest x-ray. These 20 values were averaged to produce an overall similarity/error value for each



(a) Correlation coefficients



(b) Cosine similarity



(c) Mean square error

Figure 5.8: Overall similarity values/errors for each of the 13 feature types studied

feature type.

Figure 5.8 shows the overall similarity values for each of the 13 feature types as well as the overall error value for each feature type. The height of the bars show the average (mean) similarity/error value, and the error bars represent one standard deviation above and below that mean. Note that higher values of correlation coefficient and cosine similarity are desirable, while lower values of mean square error are desirable. Our experiments indicated that out of 13 popular features types that are widely extracted to characterize images, 4 are particularly useful for this task: (1) Localized Edge Orientation Histograms (2) Haar Wavelets, (3) Gabor Filters, and (4) Steerable Filters. These were the image features that were selected for the next experiment in this work.

Image Feature	Percentage of Errors at Confidence Level						
	50%	60%	70%	80%	90%	95%	99%
Edge Orientation Histograms	50.26%	38.35%	30.82%	18.04%	9.46%	5.43%	1.23%
Gabor Filters	49.94%	39.93%	29.42%	17.69%	9.11%	4.9%	1.75%
Haar Wavelets	48.86%	36.25%	26.97%	17.34%	9.98%	5.08%	1.75%
Steerable Filters	46.94%	39.93%	29.42%	17.69%	9.11%	4.9%	1.75%

Table 5.4: Calibration results of the individual features considered in the Radiology dataset using the CP framework with ridge regression

Calibration in Multi-Regressor Fusion

In order to study the calibration of conformal predictors when the p-values are combined in a regression setting, the ridge regression based conformal predictors (detailed in Chapters 2 and 4) was used with each of the 4 image features selected in the previous section. A set of intervals, delineated by $\hat{y}_{f_i}$ in Algorithm 4, is obtained as the output for each feature f_i, $i = 1,2,3,4$. Subsequently, the $\hat{y}_{f_i}$ values of each feature are combined into a single set $\hat{y}$, which is then sorted. Now, for each interval in this sorted array, the p-values corresponding to each of the features are used to produce the combined p-value using the methods outlined in Section 5.2. Since this is a regression problem, the k-NN classifier-based combination method was not used here. The combined p-values were subsequently used to get a new set of predictions, whose 'conformity' was then studied.

Firstly, the calibration of the CP framework using ridge regression was studied with respect to each of the selected features. The results obtained are shown in Table 5.4.

Combination Method	Percentage of Errors at Confidence Level						
	50%	60%	70%	80%	90%	95%	99%
SNF	32.40%	31%	29.77%	28.02%	25.57%	23.47%	18.91%
NCA	49.56%	37.48%	28.55%	19.44%	10.16%	5.25%	1.93%
ECF	36.95%	30.65%	24.69%	18.21%	13.84%	11.03%	5.96%
MIN	73.73%	60.42%	42.73%	23.29%	10.33%	5.60%	1.93%
MAX	20.84%	16.81%	15.06%	12.61%	8.41%	4.73%	1.58%

Table 5.5: Fusion results on the Radiology dataset for the regression setting. The combination methods have been described in Section 5.2

The fusion results are presented in Table 5.5. The α regularization parameter for ridge regression was varied between 0 and 1, and the best results obtained are reproduced here. As shown in the table, the results obtained for the regression setting show more promise than what was observed for regression. Once again, a quantile combination method, Non-conformity Aggregation, performed very well and supported our claim of using these methods in real-world contexts for calibrated confidence measures in information fusion. The Standard Normal Function did not perform as well. We expect that this may have been either because the p-values did not follow a normal distribution, or because the parameters of the normal distribution chosen in this study were far from the actual parameter values. We intend to study the behavior of this method with a more varied set of parameters in future work. Also, in contrast to the classification setting, the order statistic methods, MIN and MAX, performed reasonably well, although the frequency of errors was not as expected at all the confidence levels. In summary, we conclude that the quantile combination methods are once again the best approach to combine p-values to extend conformal predictors to information fusion contexts for regression. The choice of the quantile combination method based on our study is NCA, but more empirical studies may reveal the applicability of other such methods in this context.

5.5 Summary

As the number of sensors that observe human behavior increase each day, the data generated by various modalities lay a stronger emphasis on approaches governed by information fusion. The need for reliable measures of confidence in fusion contexts has been addressed in this chapter. The CP framework provides a set of p-values for each of the classification (or regression) method involved. Subsequently, we consider the classifier (or regressor) associated with each modality as an independent statistical test, and adopt a variety of statistical techniques for combining p-values of independent tests. This methodology of obtaining a common set of p-values at the entity level, rather than at the attribute level, was validated with the multimodal person recognition problem in the classification setting, and the saliency prediction problem in the regression setting. The obtained results demonstrated that quantile methods of combining p-values (such as the Standard Normal Function and the Non-conformity Aggregation methods) provided the best calibration results, and can be considered to adopt the CP framework for information fusion.

5.6 Related Contributions

Multimodal Person Recognition

Other contributions were also made to address related problems in multimodal person recognition. In an attempt to provide an assistive face recognition system, a novel methodology for face recognition, using person-specific feature extraction and representation, was developed [260]. Distinctive facial features can take many different forms. For example, after a first encounter with a person who has a handlebar moustache, we readily recognize that person by the presence of his distinctive feature. Similarly, a person with a large black mole on her face will be remembered by first-time acquaintances by that feature. In this work, we developed a method-

ology for face recognition that detects and extracts unique features on a persons face using evolutionary learning algorithms, and then uses those features to recognize that person. The results of our research suggest that this approach can be very effective for distinguishing one persons face from other faces. For more details, please refer [260].

In yet another study, a nearest neighbors approach was proposed for face verification, and the resulting scores were combined with likelihood scores obtained from using Gaussian Mixture Models for speaker verification. Scoring methods (like minimum, maximum, average) were used for the fusion step. More details of this study can be found in [90] [224].

Saliency Prediction in Videos

Most of the existing approaches that model saliency based on visual attention are directed towards images. However, as growing numbers of videos are generated each day, there has been an equally increasing need to reliably identify appropriate regions of interest in videos. In this work, an integrated framework to learn and predict regions of interest in videos, based on human eye movements, was proposed. The eye gaze information of users is used to train a classifier to learn low-level video features from regions that attracted the visual attention of users. Such a classifier is combined with vision-based approaches to provide an integrated framework to detect salient regions in videos. The integrated approach ensures that both regions with anticipated content (top-down attention) and unanticipated content (bottom-up attention) are predicted by the proposed framework as salient. In our experiments with news videos of popular channels, the results show a significant improvement in the identification of relevant salient regions in such videos, when compared with existing approaches. For more details of this work, please refer [234] [261].

Chapter 6

ONLINE ACTIVE LEARNING USING CONFORMAL PREDICTIONS

Over the last decade, while the availability of large amounts of digital data (in the form of images, videos, speech, or text) has expanded the possibilities of solving real-world problems using computational learning techniques, active learning has emerged as a necessary component of learning frameworks to intelligently select the most relevant data samples required to build effective classifiers. In addition, annotating large amounts of data (with class labels) that are used to train the classifiers is often a very expensive process in terms of time, labor and human expertise. These factors have motivated a strong interest in newer approaches to active learning that can build effective classifiers with fewer labeled examples.

Active learning techniques primarily rely on the definition of a suitable *query function*, a function that queries each unlabeled point to decide on its appropriateness and relevance in being used to train the classifier. Query functions in existing active learning techniques often select examples that have the most uncertainty [262], least confidence [263] or maximum disagreement among a committee of classifiers [264]. Most of these existing approaches have been based on inductive inference, where a general classifier function is learnt from training examples to predict the class labels of new examples. However, there has been a growing interest over recent years in transductive inference [69], where the training examples are directly used to develop reasoning to predict the labels of new examples. In this work, we propose a Generalized Query by Transduction approach for active learning in an online (stream-based) setting using p-values obtained from a transductive inference framework introduced by Vovk et al. [38].

As mentioned in earlier sections, one of the key features of the Conformal Predictions (CP) framework [38] [54] is the calibration of the obtained confidence values in an online setting. Probabilities generated by inductive inference approaches in an online setting are often not meaningful since the model needs to be continuously updated with every new example. However, the theory behind the CP framework guarantees that the probability (or confidence) values obtained using this transductive inference framework manifest as the actual error frequencies in the online setting i.e. they are well-calibrated [56]. Further, this framework can be used with any classifier or meta-classifier (such as Support Vector Machines, k-Nearest Neighbors, Adaboost, etc). In this chapter, we propose a novel active learning approach based on the p-values generated by this transductive inference framework.

The main contributions of this work are two-fold. Firstly, we introduce the Generalized Query by Transduction (GQBT) approach for active learning using the theory of conformal predictions that can be used with any pattern classification algorithm in an online setting. Secondly, while most existing active learning approaches evaluate a single criterion (such as confidence, uncertainty or disagreement), there have been more recent efforts to combine multiple criteria (such as representativeness, informativeness and diversity by Shen et al. [265]) to select appropriate examples. We show how the proposed active learning approach can be used to combine multiple criteria for active learning. We demonstrate the improved performance of the proposed approach with commonly used datasets from the UCI Machine Learning repository [266], and apply the approach to face recognition to validate its applicability and performance in a challenging real-world problem.

In the next section, we briefly review other related techniques in active learning. The proposed GQBT approach for active learning is presented in Section 6.2,

and the illustrations of its performance on datasets from the UCI Machine Learning repository are discussed in Section 6.3. The results of the application of the proposed approach to face recognition are also subsequently discussed in Section 6.3, and the chapter concludes with potential directions of future work.

6.1 Active Learning: Background

Related Work

Several different active learning approaches have been developed over the last few years, and reviews of these approaches can be found in Baram et al. [267] and Kothari and Jain [268]. Active learning can be broadly categorized as shown in Figure 6.1. All these techniques have been developed within the scope of two distinct settings: *pool-based* and *online (stream-based)*. Pool-based active learning is further divided into Serial Query based Active Learning and Batch Mode Active Learning.

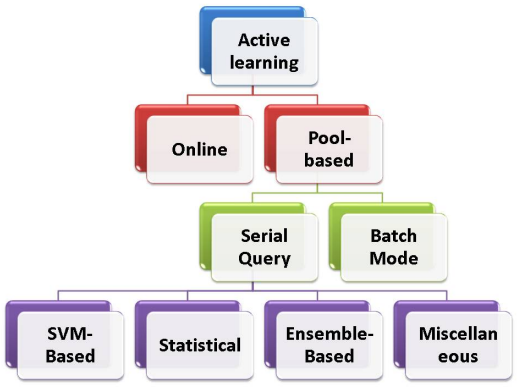


Figure 6.1: Categories of active learning

In pool-based active learning, the learner is exposed to a pool of unlabeled instances. It is assumed that the data is independent and identically distributed according to some underlying distribution $F(x)$ and the class labels y are distributed according to some conditional distribution $P(y|x)$. Given an unlabeled pool U, an

active learner has three components - (f, q, X) [269]. The first component f is the classifier which is trained on the current set of labeled instances X. The second component q is the querying function which decides which instance(s) in the unlabeled pool are to be queried next for their class label(s). The algorithm selects a data point, which is subsequently annotated, and updates itself based on the modified training set. This is continued iteratively until the entire unlabeled pool gets exhausted or some stopping criterion is met. In a serial query based system, the classifiers are updated after every single query; in a batch mode system, a batch of points is selected at once and the classifiers are updated just once (for every batch).

In contrast, in an online setting, the learner does not have access to the entire unlabeled pool at once, but encounters the points sequentially over a period of time. At each instant, the model has to decide whether to query the given point and update the hypothesis. Most existing approaches have been typically evaluated in the pool-based setting. We now present a review of online active learning approaches, followed by existing approaches for active learning by transduction, since these approaches constitute the focus of this work. For completeness, we also present a review of the more popular pool-based active learning algorithms.

Online Active Learning: Related Work

In active learning for the online setting, the Query by Committee (QBC) algorithm, originally proposed by Freund et al. [264], stands out as a commonly used approach that is naturally well-suited to this stream-based setting. In QBC, two or more experts (classifiers) are used to predict the class label of a new example. If there is disagreement between these experts (usually measured using metrics such as Kullback-Leibler divergence [59] or Jensen-Shannon divergence [270]) beyond a specified threshold, the example is queried and used to re-train the classifiers.

Several variations of the QBC have also been proposed, such as the Kernel QBC introduced by Gilad-Bachrach et al. [271] that computationally simplified the QBC algorithm. Apart from QBC, the efforts in online active learning have been scattered. Sculley [272] examined online active learning strategies based on three different classifiers for fast label-efficient spam filtering. Cesa-Bianchi et al. [273] provided regret bounds on an active learning algorithm for learning linear thresholds from an i.i.d stream of examples. Monteleoni and Kaariainen [274] developed variants of existing online active learning approaches, including Cesa-Bianchi et al.'s approach [273], and demonstrated their practical applicability in optical character recognition. Dredze and Crammer [263] proposed an online active learner for natural language processing, where the distance of a point from the margin in a large-margin classifier is combined with parameter confidence. However, all of these approaches have relied on classifiers based on inductive inference.

Active Learning by Transduction: Related Work

In comparison, there have been relatively fewer efforts towards the development of active learning techniques using transductive inference. Yu et al. [275] designed a transductive experimental methodology that selected examples from the unlabeled pool which contributed maximally to the predictions. This work was designed for a pool-based setting. Ho and Wechsler [57] used the transductive confidence machine framework (which paved way for the conformal predictions framework) and the k-NN classifier to select the examples using the *confidence* and *credibility* values generated by the framework. However, this work was also focused on the pool-based setting. As stated by Vovk [56], transductive inference is, by design, more well-suited to learning in the online setting. More recently, Ho and Wechsler [59] designed an approach similar to the proposed approach called Query by Transduc-

tion for active learning in the online setting. The proposed GQBT approach can be considered as a generalization of this approach, and the details of this approach are presented as part of the discussion in Section 6.2.

Other Active Learning Methods: A Brief Survey

Pool Based Active Learning with Serial Query

Majority of the existing active learning approaches have been applied in the pool based setting. These methods can be broadly categorized as: SVM based methods, Statistical methods, Ensemble based methods, and Other miscellaneous approaches.

SVM based methods: A sizable number of the pool based approaches are based on the Support Vector Machines (SVM) algorithm. Tong and Koller [269] designed the query function to select the unlabeled point which is closest to the SVM decision boundary in the feature space. Tong and Chang [276] applied the same concept in the image retrieval problem where in every iteration, the point that was closest to the decision boundary was returned for labeling. Osugi et al. [277] proposed a probabilistic method of active learning which decided between labeling examples near the decision boundary and exploring the input data space for unknown pockets of points. As another example, Schohn and Cohn [278] applied active learning with SVMs in the document classification task and concluded that the classifier trained through active learning often outperforms those that were trained on all the available data. Mitra et al. [279] assigned a confidence c to examples within the current decision boundary indicating whether or not they were true support vectors. Points were then queried probabilistically according to this confidence factor. Another active learning scheme using SVMs was proposed by Campbell et al.[280] where the next point to be queried was the one which minimized a predetermined risk

function. Cheng and Wang [281] used Co-SVMs in the image retrieval problem where two SVMs trained separately on color and texture features were used to classify unlabeled data - the points which were differently classified by the two SVMs were chosen to be labeled.

Statistical methods: Statistical approaches have also been extensively applied for point selection in pool-based settings. Cohn et al. [282] computed the statistically optimal way of selecting training data where the query function was based on learner variance. By choosing the point which minimized the average expected variance over all the unlabeled points, a concrete statistical basis for querying new examples was obtained. Roy and McCallum [283] adopted a sampling approach to estimate the expected reduction in error due to the labeling of a query. The future error rate was estimated by log-loss using the entropy of the posterior class distribution on a sample of the unlabeled examples. On similar lines, Holub et al. [284] attempted to minimize the expected entropy (uncertainty) of the labels of the data points in the unlabeled pool. MacKay [285] introduced information-theoretic approaches to active learning by measuring the informativeness of each data point within a Bayesian learning framework. Cohn et al. [286] described a rudimentary form of active learning which they called selective sampling. Here, the learner proceeded by examining the information already provided and then deriving a "region of uncertainty" where it believed misclassification was still possible. Ho and Wechsler [57] investigated a transductive framework to active learning where they used k nearest neighbors as the classifier. Li and Sethi [287] proposed an algorithm that identified samples that had more uncertainty associated with them, as measured by the conditional error. Tang et al. [288] used entropy based uncertainty scores to quantify the representativeness of a data point in a natural language parsing application, which was used to design the query function. Lewis and Gale [289] applied

a probabilistic framework to active learning where the most uncertain point was chosen for manual annotation.

Ensemble based methods: In ensemble based approaches, the Query by Committee (QBC) algorithm has been extensively applied. Freund et al. [290], as well as Liere and Tadepalli [291] used the disagreement measure among a committee of classifiers to select points from an unlabeled pool. McCallum and Nigam [292] modified the Query by Committee method for estimating the document density while applying active learning to the text classification problem. They also combined active learning with Expectation Maximization to take advantage of the word co-occurrence information among the documents in the unlabeled pool. Abe and Mamitsuka [293] combined QBC with boosting and bagging. The point to be queried next was the one on which the weighted majority voting by the current hypothesis had the least margin. Argamon and Dagan [294] proposed a Query by Committee algorithm in which the committee members were probabilistically selected from a distribution conditioned by the current training set. Muslea et al. [295] proposed a naive form of QBC, which they called *co-testing*, where an unlabeled point was randomly selected on which the existing views disagreed.

Other miscellaneous approaches: In other kinds of pool based approaches, Baram et al. [296] proposed a master algorithm which estimated the progress of each active learner in an ensemble during a learning session and then dynamically switched over to the best performing one at each stage. Using three active learning algorithms (Simple, Kernel Farthest First and Self-Conf) to construct an ensemble, the authors empirically established that combining them online resulted in a better performance than using any one of them. Blum and Chawla [297] developed an algorithm based on graph-cuts to learn from both labeled and unlabeled data. Nigam et al. [298] combined the Expectation Maximization (EM) algorithm with naive

Bayes classifier to learn from labeled and unlabeled text documents. Pelleg and Moore [299] proposed a mixture model approach to solve the problem of anomalous rare category identification in an unlabeled set with minimal human effort. Schein and Ungar [300] extended the A-optimality criterion to pool based active learning using Logistic Regression classifiers. Thompson et al. [301] applied the active learning framework to two non-classification tasks: semantic parsing and information extraction. They concluded that about 44% reduction in annotation cost was achieved using active learning in these complex tasks.

Clustering techniques have also been used to boost the performance of pool-based active learning [302, 303]. There have also been efforts in incorporating contextual information in active learning. Very recently, Kapoor et al. [304] incorporated *match* and *non-match* constraints in active learning for face recognition. Qi et al. [305] presented a 2D active learning scheme where sampling was done along both sample and label dimensions. The authors proposed to select sample-label pairs to minimize a multi-label Bayesian classification error bound.

Batch Mode Active Learning

While serial query based active learning has been widely applied in various problems like text classification, image retrieval, and robotics, batch mode active learning has been comparatively less explored. Brinker [306] proposed a batch mode active learning technique which ensured that the points chosen in each of the batches are highly diverse. Diversity was measured by the angles between the hyperplanes induced by the points in the batch. Hoi et al. [307] used the Fisher information matrix as a measure of model uncertainty and proposed to select a batch of points which maximally reduced the Fisher information in the classification model. The same authors have also applied the batch mode active learning concept to the prob-

lem of content based image retrieval [308, 309] and classification of medical images [310]. Guo and Schuurmans [311] proposed an optimization-based strategy for batch mode active learning, which was extended to biometrics by Chakraborty et al. [312].

We now present the proposed Generalized Query by Transduction approach for online active learning, which is derived from the CP framework.

6.2 Generalized Query by Transduction

The p-values for each of the class labels obtained using the principles of transductive inference, as outlined in the theory of conformal predictions (Chapter 2), are used to design the query function in the proposed approach. Ho and Wechsler proposed a similar approach in [59], where the query function was limited to using the top two p-values (amongst the list of p-values obtained for all the class labels). They formally defined the *closeness* between the top two p-values, $I(x_{n+1}) = p_j - p_k$, as the measure of the quality of information in an unlabeled example in the active learning process. The example is queried if $I(x_{n+1}) < \delta$, for an empirically determined threshold δ. In the proposed approach, we generalize the query function to use all (or as many as required) p-values that are obtained using the conformal predictions framework. We also call this approach *generalized* since it can be integrated into any existing classification algorithm. In addition, we show how this framework can integrate multiple criteria in the proposed query function. We illustrate the proposed approach using suitable examples, and compare the performance of our approach with Ho and Wechsler's QBT [59], along with random sampling, Query by Committee, and a Support Vector Machine (SVM) margin-based active learner in Section 6.3.

In the proposed GQBT approach, we define a matrix C which contains the ab-

solute value of the pairwise differences between all the p-values obtained from the conformal predictions framework:

$$C_{ij}(P) = \left|P_i - P_j\right| \tag{6.1}$$

where $i, j = 1, \ldots, M$ and M is the number of classes. Since this C matrix has diagonal elements as zero and is symmetric, its eigendecomposition provides a naturally useful measure with interesting properties. The largest eigen-value of C, say $\eta(C)$, assumes values that are directly proportional to the average pairwise differences between the p-values. Further, it is possible to prove that for any given set of p-values, the matrix C will always have exactly one positive eigenvalue, which we used as a measure of disagreement in this work (please refer Appendix A). When all the p-values are equal, $\eta(C)$ is trivially zero. As the pairwise differences between the p-values increase, $\eta(C)$ increases proportionately. We now show why $\eta(C)$ provides a natural measure of the extent of disagreement between the p-values, which we intend to use in the proposed approach.

The eigendecomposition of C is given by the characteristic equation:

$$|C - \lambda I| = 0 \tag{6.2}$$

where $|.|$ is the matrix determinant. When the pairwise differences are multiplied by a constant factor, say d, the new C, say C^*, is equal to dC. The characteristic equation for C^* is given by:

$$|C^* - \lambda^* I| = 0 \tag{6.3}$$

where λ^* are the eigenvalues of C^*. Substituting $C^* = dC$,

$$|dC - \lambda^* I| = 0 \Rightarrow \left|d(C - \frac{\lambda^*}{d} I)\right| = 0 \tag{6.4}$$

$$\Rightarrow |dI| \left|C - \frac{\lambda^*}{d} I\right| = 0 \tag{6.5}$$

Since $|dI| \neq 0$,

$$\Rightarrow \left| C - \frac{\lambda^*}{d} I \right| = 0 \tag{6.6}$$

Comparing Equations 6.6 and Equation 6.2,

$$\lambda = \frac{\lambda^*}{d} \tag{6.7}$$

that is, the eigenvalues λ^* are also multiplied by the same constant factor d. For another C matrix, say $\hat{C}$, whose average pairwise difference lies between the original average pairwise difference in C and that in C^*, the corresponding eigenvalues $\hat{\lambda}$ will lie between λ and λ^*. We exploit this ordering of eigenvalues as a natural measure of the extent of disagreement among the p-values obtained.

Since p-values assume values in the interval $[0,1]$, the largest eigenvalue, $\eta(C)$, tends to have low numeric values. For convenience of implementation, we compute the inverse of C, and use the largest eigenvalue of C^{-1} in our work. Since $\eta(C^{-1})$ is *inversely* proportional to the average difference between the p-values, we accordingly factor this in the design of our query condition. The proposed GQBT approach is presented in Algorithm 5.

Almost all online active learning algorithms rely on empirically obtained thresholds to decide if an unlabeled example needs to be queried. For example, Ho and Wechsler [59] start the active learner with a random threshold and update the threshold with the average of the previous n p-value differences, when a sequence of n examples are not queried in succession. In contrast, in this approach, the largest eigenvalue has a straightforward connotation that can be exploited. The selection threshold δ is initialized to the largest eigenvalue of the C^{-1} matrix that is constructed assuming the pairwise differences between the p-values are equal to a unit percentage (i.e. 0.01) each. Similar to what was proved in Equation 6.7, the eigenvalues for C^{-1} are divided by a factor of d, when C is multiplied by d. Hence,

Algorithm 5 Generalized Query by Transduction for Online Active Learning

Require: Training set $T = \{(x_1, y_1), ..., (x_n, y_n)\}$, classifier Ξ, selection threshold δ, stopping threshold γ, number of classes M, number of queried points p, budget constraint β (maximum number of points that can be queried)

initialize $p \leftarrow 0$
repeat
  Get new unlabeled example x_{n+1}.
  for all class labels, y_i, where $i = 1, \ldots, M$ **do**
    Assign label y_i to x_{n+1}.
    Update the classifier Ξ, with $T \cup \{x_{n+1}, y_i\}$.
    Compute non-conformity measure value, $\alpha_{n+1}^{y_i}$ to compute the p-value, P_i, w.r.t. class y_i (Equation 2.7) using the conformal predictions framework.
  end for
  Construct the matrix C, such that $C_{ij}(P) = \left|P_i - P_j\right|$ (Equation 6.1).
  Compute $\eta(C^{-1})$ as the largest eigenvalue of C^{-1}.
  if $\eta(C^{-1}) > \delta$ **then**
    Add x_{n+1} to training set i.e. $T \leftarrow T \cup \{x_{n+1}, y_c\}$, where y_c is the correct label for x_{n+1}.
    $p \leftarrow p + 1$.
  end if
until $\eta(C^{-1}) > \gamma$ or $p < \beta$

when the pairwise differences are equal to 0.02 each, the largest eigenvalue of the corresponding C^{-1} matrix is now equal to $\frac{\delta}{2}$. To apply this in the algorithm, if no examples are selected after, say r, examples are observed, the selection threshold is changed to: $\delta \leftarrow \frac{\delta}{2}$, thus allowing for a more accommodative threshold. Depending on the dataset under consideration, this can progressively be continued at periodic intervals to $\delta \leftarrow \frac{\delta}{3}, \delta \leftarrow \frac{\delta}{4}$, and so on, as may be required in a particular setting. This provides for an automatic methodology to set (and modify) threshold values, where the query condition becomes lenient with time.

We use Support Vector Machines (SVM) as the classifier in this work for a few reasons. Firstly, there have been several active learning techniques in the recent past that have used the margin distance in a SVM to query examples in active learning [313] [59], leading to the popularity of SVMs in active learning. Secondly, there

have been recent efforts to develop incremental SVMs for an online setting [314] to train newer examples into an existing SVM model. One of the primary limitations of the proposed approach (or any transductive inference approach, for that matter) is the computational overhead in Steps 5-7 in Algorithm 5 for each class label. The use of incremental SVMs substantially offsets this limitation. Thirdly, the Lagrange multipliers obtained while training a SVM are a straightforward choice to consider as non-conformity scores, as pointed out by Vovk et al. [38]. The Lagrange multipliers, $\alpha_i, i = 1,\ldots,n$, are computed while maximizing the dual formulation in the soft margin SVM:

$$Q(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) + \sum_{i=1}^{n}\alpha_i \tag{6.8}$$

subject to constraints $\sum_{i=1}^{n}\alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C, i = 1,\ldots,n$, and $K(.)$ is the kernel function. The Lagrange multipliers' values are zero for examples outside the margin, and lie between 0 and C for examples on and within the margin, thereby providing a natural monotonic measure of *non-conformity* w.r.t. the corresponding class.

To extend the definition of the non-conformity measure to a multi-class SVM, we adopt the 'one-against-the-rest' approach, and define a binary SVM for each of the class labels. The non-conformity measure is then computed as suggested by Vovk et al. [38] (pg 59) using the non-conformity measures computed from each of the individual SVMs, α_i^y:

$$A_i^{y_p} = \lambda\,\alpha_i^{y=y_p} + \frac{1-\lambda}{M-1}\sum_{y\neq y_p}\alpha_i^y \tag{6.9}$$

where M is the number of classes, $p = 1,\ldots,M$, and λ is a parameter that is empirically determined. Equation 6.9 states that the non-conformity measure for a data instance, x_i, in the multi-class SVM is a normalized sum of the non-conformity

values of x_i w.r.t $y = +1$ class in the binary SVM constructed for class y_p, and the $y = -1$ class in all the other binary SVMs constructed for the multi-class model.

Why Generalized QBT?

Before we present the experimental results, we show how the proposed GQBT is a generalization of the QBT approach proposed by Ho and Wechsler [59]. Ho and Wechsler define the quality of information of a new data example as $I(x_{n+1}) = p_j - p_k$, where p_j and p_k are the highest 2 p-values obtained using the conformal predictions framework. We define the quality of information using the largest eigenvalue of the matrix C containing the pairwise differences between all p-values. In a binary classification problem (or if only the top 2 p-values are used in a multi-class setting), our approach becomes the same as Ho and Wechsler's. This is because C is now given by:

$$\begin{bmatrix} 0 & |p_1 - p_2| \\ |p_1 - p_2| & 0 \end{bmatrix}$$

whose largest eigenvalue is $|p_1 - p_2|$ itself, which is the measure used by Ho and Wechsler. However, the progressive choice of selection threshold values in our approach (as $\delta, \frac{\delta}{2}$, etc. detailed earlier) performs better than the empirical choice of thresholds in Ho and Wechsler's approach. This is illustrated in Figure 6.2, which shows how the proposed GQBT approach has a lower label complexity i.e. it achieves the highest accuracy by querying much fewer points than Ho and Wechsler's approach.

Combining Multiple Criteria for Active Learning

It may often be essential to combine multiple criteria to decide if a particular unlabeled example needs to be queried for its true label, and included in the training set, and there have been recent efforts in this direction [265]. In our work, for example,

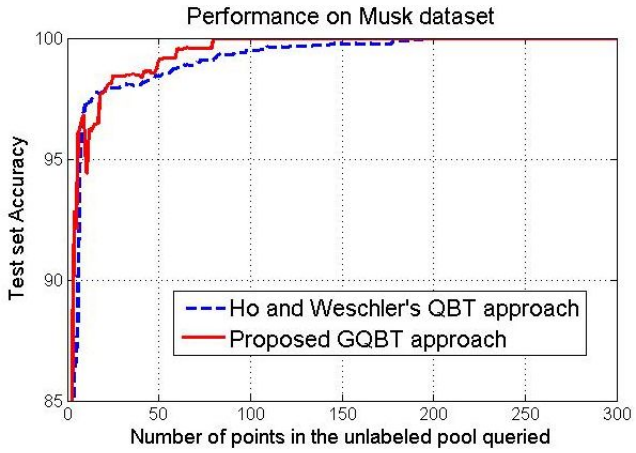


Figure 6.2: Comparison of the proposed GQBT approach with Ho and Wechsler's QBT approach on the Musk dataset from the UCI Machine Learning repository. Note that our approach reaches the peak accuracy by querying $\approx$ 80 examples, while the latter needs $\approx$ 160 examples.

in addition to the Lagrange multipliers (whose values are closely related to the distance of the example from the SVM margin), it may be useful to consider another non-conformity measure that estimates the density of examples in the neighborhood of a given unlabeled example. This can be defined using the k-NN classifier (a non-parametric density estimator), as stated earlier in Equation 2.6 in Chapter 2. Evidently, the theory of conformal predictions can also be used with this measure to obtain another set of p-values. We use results from statistical hypothesis testing to combine these p-values. Given that the p-value is a uniformly distributed random variable on the interval $[0,1]$, the combined significance level or p-value of n individual p-values can be given as [6]:

$$k \sum_{i=0}^{n-1} \frac{(-\ln k)^i}{i!} \tag{6.10}$$

where $k = (p_1 \times p_2 \times p_3 \ldots \times p_n)$, the product of the given set of p-values. While we use this approach in our work, there are other methods in hypothesis testing to combine p-values [205], which can be used too. Please refer to Chapter 5 of this

dissertation for details on more such methods. A similar approach was also used in [58] for head pose classification.

Figure 6.3 shows the improvement in performance obtained (on the same dataset as in Figure 6.2) by combining the p-values obtained using the non-conformity measures computed from the SVM and the k-NN classifier. Note the reduction in label complexity obtained by combining the p-values from the two non-conformity measures discussed in Section 6.2. The proposed approach needs only ≈ 50 examples to reach the peak accuracy.

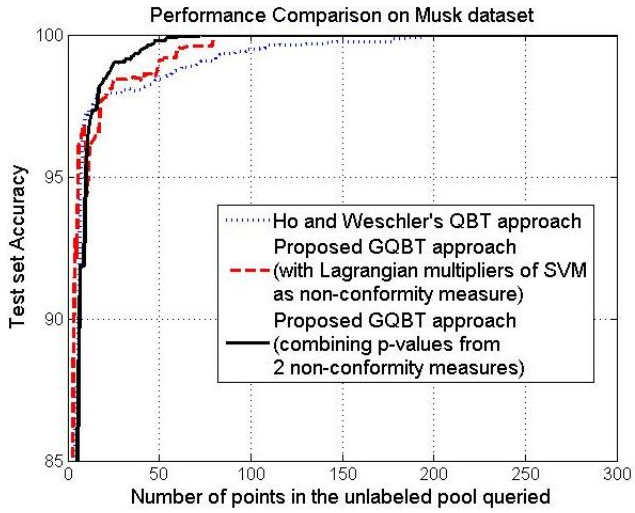


Figure 6.3: Performance comparison on the Musk dataset (as in Figure 6.2)

6.3 Experimental Results

We compared the performance of the proposed GQBT approach with three other online active learning algorithms together with random sampling. The methods are briefly outlined below:

- *Random Sampling:* In this method, when a new example arrives, we randomly decide whether to query this point for its class label or not, i.e. each example is queried with a probability of 0.5.

- *Margin based SVM:* An SVM classifier is constructed from the given set of training instances. For an unlabeled example x_{n+1}, its decision value $f(x) = w.\phi(x) + b$ is computed and if it is below a certain threshold, the point is queried. If a certain number of unlabeled points are not queried in succession, the threshold is updated as the average of the SVM decision values of the unqueried examples.

- *Query by Committee:* A committee consisting of two classifiers, SVM and k-NN (with $k = 10$), was used. For a given unlabeled example, the SVM output values are converted into probabilities using Platt's method [315]. For k-NN, the class probability for the unlabeled example is defined as the fraction of the number of points of a given class occurring in its k nearest neighbors. Once we have the probability values from the two classifiers, we compute the Kullback Leibler divergence between these two sets. A high divergence implies that the point is informative and should be queried. The threshold for the KL divergence value was updated as described for the margin based SVM.

- *Query by Transduction:* This is the method proposed by Ho and Wechsler [59] as described previously.

We selected five datasets (with different number of classes, dimensions and instances) from the UCI Machine Learning repository [71] to test the generalizability of the proposed approach. The datasets and their details are listed in Table 6.1. An equal number of examples from each class was used in the initial training set. For example, for the Breast Cancer dataset, 5 examples from each class were used to form the initial training set of 10 examples. For each of the datasets, the initial training, testing and unlabeled pools were randomly partitioned three different

Dataset	Classes	Size	Dimensions	Initial training set	Size of unlabeled pool	Size of test set
Breast Cancer	2	569	30	10	259	300
Musk	2	1000	166	2	498	500
Wine	3	178	13	3	88	87
Waveform	3	5000	21	15	2485	2500
Image Segmentation	7	2310	19	35	175	2100

Table 6.1: Datasets from the UCI Machine Learning repository used in our experiments

times and the results were averaged from these 3 runs. Further, in each of the runs, the unlabeled pool was randomly permuted 10 different times to remove any bias on the order in which the points are observed, and the results of these 10 trials were averaged for each run. A polynomial kernel was found to be the most well-suited for all the datasets, as established by the peak accuracies achieved in our results.

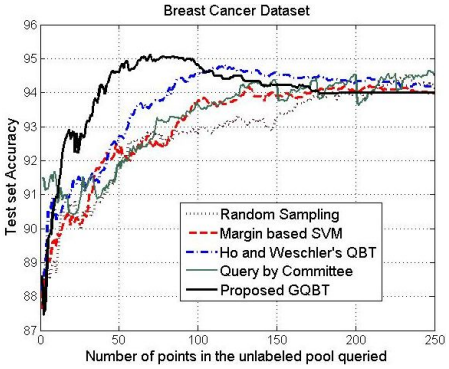
The results of our experiments are presented in Figure 6.4 and Table 6.2. In each of these experiments, the formulation of the proposed GQBT approach where the non-conformity measures from the SVM and the k-NN are combined (as in Section 6.2) was used. Table 6.2 shows the label complexity (the percentage of the unlabeled pool that was queried to reach the peak accuracy in the active learning process) of each of the methods. The results are self-explanatory, and demonstrate the improvement in performance gained using the proposed approach. Label complexity is defined as the percentage of the unlabeled pool that is queried to reach the peak accuracy in the active learning process. In Table 6.2, note the low label complexities of the proposed approach in all the cases. Also, note that the label complexities for the other methods on datasets like Waveform and Image Segmentation are very high although the accuracy did increase at a reasonable rate in the

active learning process in Figure 6.4. This only implies that these methods reached their peak accuracy when the unlabeled pool was almost exhausted. Note that in the Musk dataset, the results started with an accuracy of $\approx 70\%$, but since all methods had similar initial accuracies, the graph is shown from 85% accuracy onwards, where the differences in performance are clearly seen.

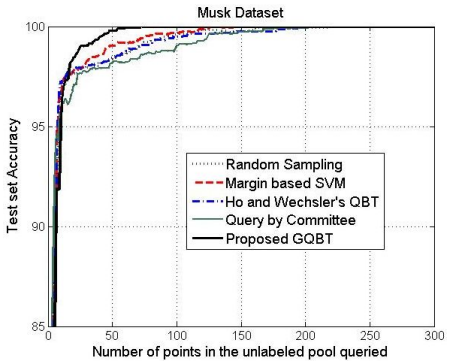
Dataset	Random Sampling	Margin-based SVM	Query by Committee	Ho-Wechsler's Initial QBT	Proposed GQBT
Breast Cancer	92.8%	83.6%	80%	46.8%	28%
Musk	77%	55%	72.33%	86.67%	24.33%
Wine	87.5%	78.75%	97.5%	47.5%	35%
Waveform	99.6%	100%	98.2%	98.6%	89.2%
Image Segmentation	100%	100%	100%	98.18%	66.06%

Table 6.2: Label complexities of each of the methods for all the datasets. Label complexity is defined as the percentage of the unlabeled pool that is queried to reach the peak accuracy in the active learning process

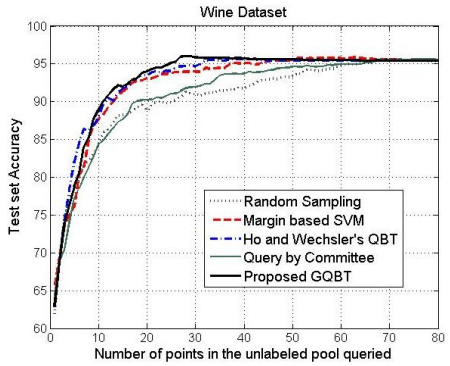
To evaluate the performance of the approach on a challenging real-world problem, we carried out experiments on face recognition from video, where the high redundancy between frames in a video requires an active learning approach. We used the VidTIMIT biometrics dataset [316] (described in Chapter 2), of which we used the video recordings of 25 subjects reciting short sentences. Each of the videos are sliced and stored as JPEG images of resolution 512 by 384, on which automated face cropping was performed to crop out the face regions. To extract the facial features, block based discrete cosine transform (DCT) was used (similar to [92]). Each image was subdivided into 8 by 8 non-overlapping blocks, and the DCT coefficients of each block were then ordered according to the zigzag scan pattern. The DC co-efficient was discarded for illumination normalization, and the first 10 AC co-efficients of each block were selected to form compact local feature vectors.



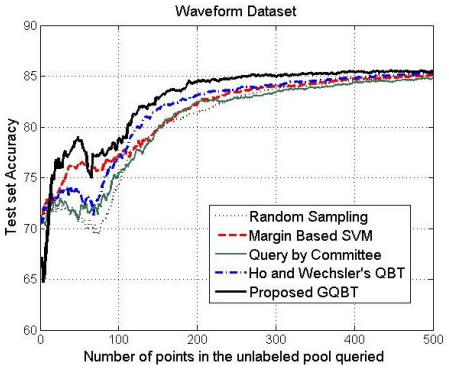
(a) Breast Cancer dataset



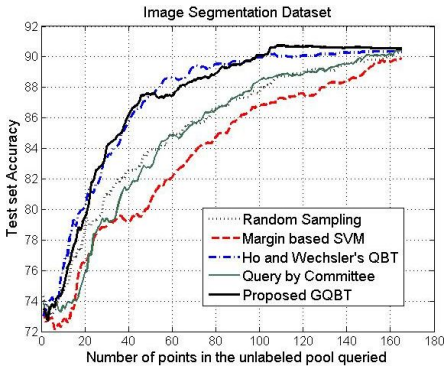
(b) Musk dataset



(c) Wine dataset



(d) Waveform dataset



(e) Image Segmentation dataset

Figure 6.4: Results with datasets from the UCI Machine Learning repository

Each local feature vector was normalized to unit norm. Concatenating the features from the individual blocks yielded the global feature vector for the entire image. The cropped face image had a resolution of 128 by 128 and thus the dimensionality of the extracted feature vector was 2560. Principal Component Analysis (PCA) was then applied to reduce the dimension to 100, retaining about 99% of the variance.

50 images of each subject were randomly picked, and divided into the initial training set (10), unlabeled pool (20) and the test set (20). A polynomial kernel was used for the SVM classifier. Similar to the previous set of experiments, the unlabeled pool was randomly permuted 3 different times to remove any bias on the order in which the points are observed, and the results of these 3 trials were averaged. Figure 6.5 shows the results of our experiments. As shown, the proposed GQBT once again demonstrated a significantly improved performance over the other approaches. Note that the GQBT approach led to a significantly higher peak accuracy, and had a lower label complexity of 58.8% to reach the peak accuracy. Label complexities of the other methods: Ho and Wechsler's QBT - 98.2%; Query by Committee - 100%; Margin-based SVM - 89%; Random sampling - 99.6%.

6.4 Summary

Transductive inference has gained popularity in recent years as a means to develop pattern classification approaches that address the specific issue of predicting the class label of a given data point, instead of the more general problem of inferring the ideal classifier function. A Generalized Query by Transduction (GQBT) approach for active learning in the online setting, based on the theory of conformal predictions, has been presented in this work. The proposed GQBT approach can be used along with any existing pattern classification algorithm, and can also be used to combine multiple criteria in selecting an unlabeled example appropriately in the

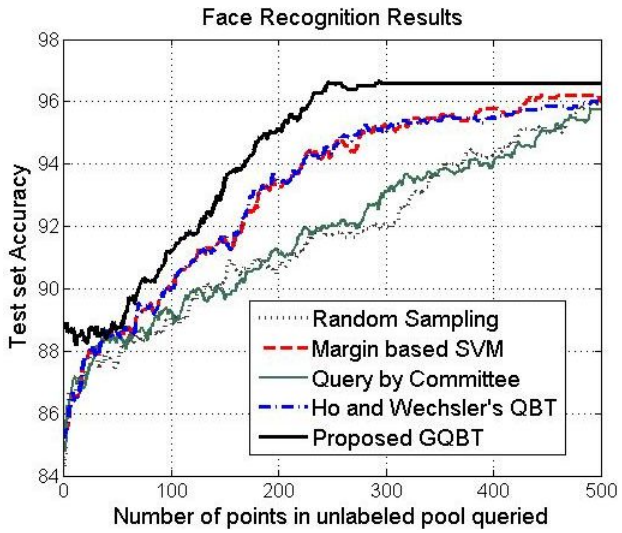


Figure 6.5: Results obtained for GQBT on the VidTIMIT dataset

active learning process. The results of our experiments with different datasets from the UCI Machine Learning repository demonstrate high promise in the proposed approach, with significantly lower label complexities than other existing online active learning approaches. The GQBT approach was also evaluated on person recognition using videos from the VidTIMIT dataset, and showed superior performance in this set of experiments too, supporting the potential of applicability of the proposed approach in real-world problems.

One of the major limitations of this approach, as mentioned earlier, is the computational overhead of transductive inference at each step. With recent advances in incremental classifiers, this limitation can be overcome to a large extent. In future work, we will study the performance of using the inductive flavor of the Conformal predictions framework [63] to offset the computational overhead. We also plan to study other approaches of combining p-values and their influence on the performance of the approach. Further, we intend to study and identify appropriate stopping criteria for the proposed active learning framework.

6.5 Related Contributions

A related contribution on Batch Mode Active Learning for biometrics has also been made as part of this work. In biometric applications like face recognition, real world data is usually generated in batches such as frames of video in a capture session. The captured data has high redundancy and it is a significant challenge to select the most promising instances from this superfluous set for training a classifier. In this work, a novel batch mode active learning scheme has been formulated, where the instance selection is based on numerical optimization of an objective function, which can be adapted to suit the requirements of a particular application. The results obtained on the person recognition problem certify the potential of this method in being used for real world biometric recognition problems. For more details of this work, please refer [312] [317].

Chapter 7

CONCLUSIONS AND FUTURE DIRECTIONS

The increasing application of machine learning, data mining and pattern recognition methods in data-rich fields across various domains has made the reliable estimation of confidence a universally relevant field of research. As pointed out in Chapter 1, there have been extensive efforts to resolve both aleatory as well as epistemic types of uncertainty using techniques such as probability theory, Monte Carlo simulations, evidence aggregation theories, fuzzy logic and statistical hypothesis testing methods. There have been critical debates on the appropriate forms of representing uncertainty using methods including probability values, confidence intervals, credible intervals and gamesman intervals. However, the need for a refreshingly different approach towards the estimation of uncertainty based on practical application rather than asymptotic guarantees, has remained accentuated over the years.

As described in Chapter 2, recent efforts towards the estimation of confidence in machine learning algorithms have resulted in the development of an emerging game-theoretic approach for hedging predictions. Vovk, Shafer and Gammerman [38] developed the Conformal Predictions framework based on the relationship derived between transductive inference and the Kolmogorov complexity of an identically independently distributed (i.i.d.) sequence of data instances. This framework, which can be used with any machine learning classification or regression algorithm, outputs prediction regions based on a user-defined confidence level. This approach has desirable theoretical guarantees including calibration of errors in an online setting. This property of calibration of errors, with respect to a user-defined confidence level, is termed ‘validity’ in the framework. However, theoretical valid-

ity does not guarantee practical usefulness, and the applicability of this promising framework in real-world contexts as an effective and efficient real-time reasoning tool requires several significant computational challenges that need to be addressed.

Hence, the objective of this dissertation was to design and develop learning methodologies and pattern recognition models that provide conformal predictions for decision-making in realistic settings. We specifically focused on the development of Conformal Prediction methods for machine learning based predictive models in multimedia pattern recognition, with applications in healthcare (cardiac decision support and radiology) and assistive technology systems (for individuals with visual impairments). We note, however, that the research outcomes of this work are fundamental by their impact, and the solutions developed as part of this dissertation for these application domains are pertinent to a broader audience, including many other related fields of healthcare and medicine, risk-sensitive financial management models and security applications.

The fundamental intellectual merit of this work lies in the transformational nature of the application of conformal predictive models, which can provide error guarantees in risk-sensitive applications across various fields. This work demonstrates how a sound computational framework based on fundamental theories can translate to practical usefulness in many different domains. The intellectual merit lies not only in strong contributions in pattern recognition and machine learning, but also opens up new research directions at the intersections of these disciplines, and in healthcare informatics, assistive technologies, disability studies, and cognitive decision sciences.

7.1 Summary of Contributions

The key contributions made in this dissertation are summarized in Figure 7.1. While this figure mentions only the classification setting, this work has proposed these contributions from a regression perspective too. As listed in Chapter 1, this work

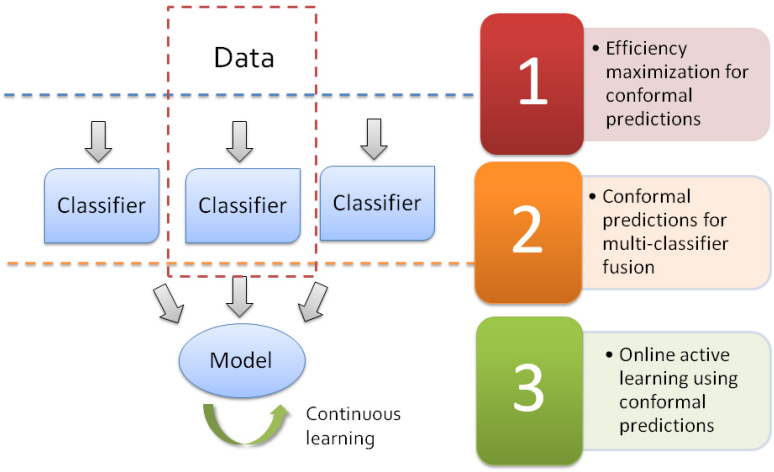


Figure 7.1: Summary of the contributions made in this work

makes three specific contributions that aim to make the CP framework practically useful in real-world multimedia pattern recognition problems. Figure 7.1 illustrates that these contributions can be viewed as steps towards a holistic learning system.

1. To develop methods that can compute efficient conformal predictions for a given classifier.
2. To study and propose solutions to retain the calibration property when conformal predictors are applied to multi-classifier fusion.
3. To extend the framework to continuous online learning, where the measures of confidence computed by the framework are used for online active learning.

These contributions have been validated in classification settings using the problems of risk prediction in cardiac decision support and multimodal person recognition, and in regression settings using head pose estimation and saliency prediction in images. During this work, other contributions related to addressing the problems of each of these applications domains (such as a generalizable framework for supervised manifold learning and a clinically relevant inter-patient distance metric) have been proposed.

7.2 Summary of Outcomes

The specific outcomes of this dissertation have been listed below:

- *Dissemination:* The various aspects of the contributions in this work have resulted in a total of 23 peer-reviewed conference publications, 5 journal publications/book chapters and 1 provisional patent. The dissemination venues include IEEE Computer Vision and Pattern Recognition (CVPR), Neural Information Processing Systems (NIPS), IEEE International Conference on Computer Vision (ICCV), International Conference on Pattern Recognition (ICPR), ACM Multimedia, and the Springer Lecture Notes on Computer Science series.
- *Datasets and Code:* While existing publicly available datasets were used for multimodal person recognition, datasets were created for each of the other three application domains: head pose estimation, cardiac decision support and saliency prediction in radiology. The FacePix dataset for head pose estimation has been made publicly available[1]. A Matlab toolbox with inductive and transductive approaches to Conformal Predictions for classification and

[1] http://www.facepix.org

regression has been implemented. Efforts are being undertaken to make this toolbox publicly available within the next few months.

- *Funding Opportunities:* The work in this dissertation resulted in submission of grant proposals to the American Heart Association and National Institute of Health (on 'Advanced Computational Techniques in Cardiovascular Disease: Individualized Prediction of Outcomes Following Coronary Stenting Using Clinical Machine Learning') and the National Science Foundation (on 'Conformal Predictions in Healthcare and Assistive Technologies').
- *Related Theses:* In association with the work in this dissertation, the problems in the application domains have resulted in two completed Masters theses (one on saliency prediction in videos, and another on clinical machine learning in interventional cardiology), and a continuing PhD thesis (on batch mode active learning for biometrics).

7.3 Future Work

The contributions of this dissertation have shown the promise and merit of using conformal predictors for reliable estimation of confidence in various multimedia pattern recognition problems. As shown in Figure 7.2, while machine learning based predictive models are used increasingly in different applications, there still exists a gap of trust between end users and the predictive models (physicians and predictive patient models, for example). The contributions in this work attempt to narrow this gap by providing the user with the ability to control a level of confidence that is obtained from a system.

The framework, however, is in its nascent stages, and is slowly being absorbed into related bodies of work. The possibilities of future work are numerous, and a few sample directions are presented in this section.

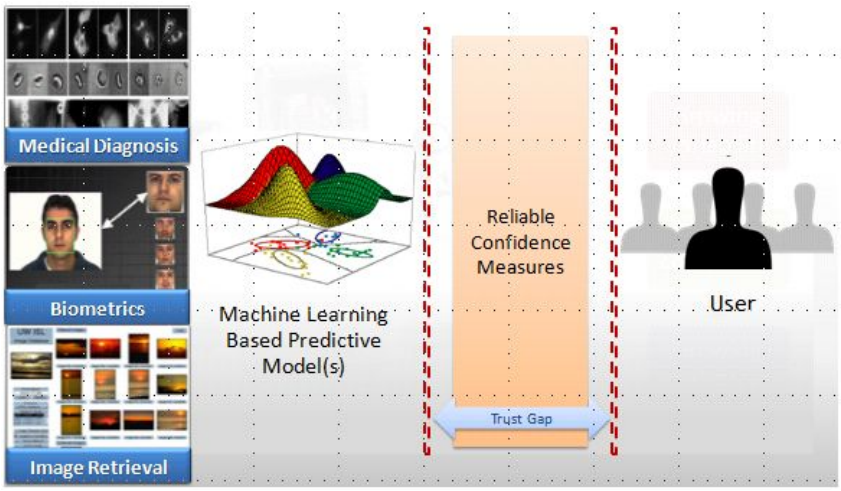


Figure 7.2: A high-level view of this work

Efficiency Maximization

- An alternate formulation for Multiple Kernel Learning in efficiency maximization was proposed for the k-NN classifier in Chapter 3. In this alternate formulation, the kernel function was incorporated from the start of the derivation, rather than the other formulation where data was projected to convert the SVM-LDA problem to a SVM problem, and kernel learning was subsequently performed. It is expected that the alternate formulation will provide better results for minimizing multiple predictions (or maximizing efficiency). In future work, we will study the effectiveness of this alternate formulation for the cardiac decision support problem, and also investigate the connections (or equivalence) between the two formulations.

- The kernel learning formulation for the classification context in Chapter 3 was aimed at ensuring that the non-conformity score of a test data instance with respect to the 'correct' class label is low. However, the discussion of the conceptual framework was tailored towards the non-conformity score of the

k-NN classifier. While this is a fairly desirable criterion that can generalize to many other classifiers, we plan to revisit the conceptual framework of this portion of the work from the viewpoint of generalizing to non-conformity scores commonly used for other classifiers.

- In the regression setting (Chapter 4), a supervised manifold learning approach was used as a form of metric learning for maximizing efficiency. However, it is possible that metric learning with a specific objective function tailored to increase efficiency can be designed too. Also, when kernel ridge regression is used (or an equivalent kernel version of any regression technique), a suitable kernel function can be learnt (similar to what was proposed for classification in Chapter 3). Both these directions are potentially interesting and valuable for future work.

- The CP framework for regression has so far been most efficiently defined for ridge regression, and this algorithm was used in this work to conceptualize the methodology for efficiency maximization. There is a need to define suitable variants of the approach for other regression methods too.

Information Fusion

- In combining p-values from multiple hypothesis tests, this dissertation (Chapter 5) considers the individual modalities (or individual features from a single modality) to be independent. Quantile combination methods and order statistic methods have typically been studied for combining p-values from independent tests. However, it is possible that the individual hypothesis tests (corresponding to each modality or feature) are dependent, necessitating a different approach to combine the p-values from the individual tests or modalities. In real-world scenarios, it is more likely that the different data sources have a

certain extent of dependency. Several methods described to combine p-values of dependent tests have been discussed by Pesarin in [204], and using these methods to combine p-values in fusion settings and studying calibration is a very necessary future step of this work.

- While the method proposed in this work to combine p-values from attribute-level classifiers (or regressors) generates p-values at an entity level, it would have to be investigated to understand if there are any relationships between the obtained values. For example, in the multimodal person recognition problem, if the user would like to understand which of the modalities resulted in errors (so that appropriate corrections can be made, possibly in the form of an additional training phase for that modality), it would be essential to capture the relationships between the p-values, and identify appropriate contributing factors to the entity-level values. Such a study can provide insights that may be potentially useful in several ways, including serving as a criterion for active learning or transfer learning techniques.
- Further, in this dissertation, we studied only the calibration properties of multiple classifier/regressor fusion using the p-values obtained from the CP framework. However, for practical usage - as mentioned earlier - it is essential to consider the efficiency too. Hence, a logical extension of this work would be to combine the methodology proposed for efficiency maximization with the fusion approach for more appreciable results in real-world problems.

Active Learning

- In the Generalized Query by Transduction (GQBT) approach (Chapter 6), the largest eigenvalue of the p-value difference matrix provides a convenient measure of uncertainty to decide if a particular data instance needs to be

included for further training. The results showed substantial improvements over other related approaches. However, it would be necessary to investigate the connections between this discrepancy measure and other measures such as Shannon entropy or Jensen divergence. Further, it is also possible to define a similar p-value difference matrix for Inductive Conformal Predictors. Hence, an equivalent Generalized Query by Induction method can be proposed along the same lines, and such a method will have significantly lower computational costs.

- An important question, often neglected in existing active learning methods, is the choice of a suitable stopping criteria that can indicate when further training is not required. As observed in some of our results, the accuracy of a classifier can sometimes reach a peak value, and then start falling down as more training data examples are used. Overfitting may be one possible reason to explain this observation. Nevertheless, this observation vindicates the need for defining a suitable stopping criterion in the proposed GQBT approach.

Other Possible Directions

Limitations specific to each of the contributions in this dissertation were stated in the above paragraphs. However, there are other possible directions of future work that can impact all of the contributions. Pointers to these directions are briefly described below.

- *Inductive Conformal Predictors:* An inherent limitation of transductive approaches is the computational overhead involved in re-training an entire set of data instances, whenever a single new test data instance is encountered. This limitation is a hindrance for adoption of this work in real-time application contexts. In order to address this issue, we plan to study the performance

of each of the contributions in this dissertation with the Inductive Conformal Predictions framework [63] [48]. In this inductive approach, a subset of training data is randomly set aside as a 'calibration' set. When a new test data instance arrives, the p-values in the CP framework are computed with respect to the data instances in the calibration set, thus avoiding the need for re-training the model. Our preliminary experiments in this direction suggested that this approach provides significant computational benefits, and we intend to follow this line of work in the near future.

- *Mondrian Conformal Predictors:* Another issue that is prominent especially in healthcare related applications is that data often is imbalanced in terms of the class labels. In one of the related contributions of this work [78], we performed a study that showed the effectiveness of Synthetic Minority Oversample Technique (SMOTE) as a data processing method to handle such class imbalance. However, when the CP framework is applied to such a scenario, the calibration property may not be as meaningful, since all the errors may be concentrated in a specific class (or a localized subset of classes). To address this issue, Vovk et al. [38] proposed a modified version of the CP framework called Mondrian conformal predictors, which from one perspective, is a generalized view of the transductive and inductive variants of the framework. Mondrian conformal predictors can guarantee *conditional validity*, i.e. calibration within each of the class labels in a given problem. This is very desirable in risk-sensitive applications, and we plan to study the usefulness of this approach in our future work. We will also work on maximizing efficiency in the context of the Mondrian framework.

- *Transfer Learning using Conformal Predictors:* The problem of transfer learn-

ing, where information gained in one learning task is used to improve performance in another related task, is an important new area of research. Several of the existing approaches use measures such as those based on entropy [318] or conditional independence [319]. Naturally, the measures obtained from the CP framework (either the non-conformity scores, p-values or the discrepancy measure obtained in the GQBT approach) can be potentially used in transfer learning contexts.

Application Perspectives

The problems addressed in each of the applications in this dissertation are also potential avenues of future research. While pointers to these opportunities were presented in the respective chapters, a few samples are summarized below.

- *Cardiac Decision Support:* Advances in medical technologies have resulted in a tremendous increase in the quantity of available information in terms of patient records, diagnostic tests, genomics, treatments, etc. Further, there has been a paradigm shift in the field from evidence based medicine to personalized medicine. This shift only reiterates the need to reduce uncertainty in predicting patient outcomes as a function of treatment [65]. Using efficient and valid conformal predictors in various kinds of personalized predictive models in biomedicine presents a potential direction of this work. Also, while efficiency of these predictors was maximized using kernel learning in this work, it is possible to design clinically relevant kernels that use knowledge bases such as ontologies together with data [134]. This presents even more interesting challenges for future work.
- *Multimodal Biometrics:* As security and surveillance systems assume greater roles in crime prevention, the need to provide reliable estimates of the identi-

ties of persons of interest in real-time remains a significant challenge. The extension of the CP framework to information fusion contexts is a valuable first step in this direction, but a lot more needs to be done for such contributions to be applied in on-field scenarios. While inductive conformal predictors can be adopted to provide real-time performance, the validity of the calibration properties of the framework on large-scale datasets remains in question, and is a key issue to be studied. In addition, this work did not consider temporal characteristics of speech and video signals while extending the CP framework. Exploring the properties of the framework for temporal sequences is another important problem to be investigated.

- *Person-Independent Head Pose Estimation:* While real-time head pose estimation technologies have reached stages of commercialization (for example, FaceAPI [2]) in recent years, fundamental research issues lie unsolved in capturing low-dimensional representations of data using manifold-based approaches. Identifying the intrinsic dimensionality of data with geometric relationships remains a challenge. Further, since the objective of this work was to equip a wearable Social Interaction Assistant with real-time pose estimation, computer vision problems related to wearable systems (such as decimating redundant frames or deblurring frames in a video captured by a moving person) constitute issues that need to be addressed in this regard.

- *Saliency Prediction:* Using eye gaze of experts to train a machine learning model that can predict saliency in data such as medical images is a novel concept that is still in its early stages. One of the lessons learnt in this work was the need to develop predictive models that can learn with very little data,

[2] http://www.seeingmachines.com

since medical images with eye gaze information is not very easily available. While there are widespread efforts in building models that learn from large datasets in the last few years, there is also a need to develop models that can learn from very small datasets. This is a fundamental problem that can have impact in many fields. Additionally, saliency is a very personal and subjective concept - building models that can be adapted over time to each individual user presents a problem, which when solved can have great impact.

BIBLIOGRAPHY

[1] T. Melluish, C. Saunders, I. Nouretdinov, and V. Vovk, "Comparing the bayes and typicalness frameworks," *In Proceedings of the 12th European Conference on Machine Learning (ECML-2001*, vol. 2167, pp. 360—371, 2001.

[2] S. Prabhakar and A. K. Jain, "Decision-level fusion in fingerprint verification," *Pattern Recognition*, vol. 35, pp. 861—874, 2001.

[3] L. Yang, "Distance metric learning: A comprehensive survey," Department of Computer Science and Engineering, Michigan State University, Tech. Rep., May 2006.

[4] M. Balasubramanian and E. Schwartz, "The Isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, 2002.

[5] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems?" *Pattern Recognition*, vol. 38, no. 12, pp. 2285, 2270, Dec. 2005.

[6] L. Jost, "Combining significance levels from multiple experiments or analyses," http://www.loujost.com/Statistics and Physics/StatsArticlesIndex.htm, last Accessed: Jun 18, 2009.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[8] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[9] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 6, pp. 989–1005, 2009.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, Oct. 2007.

[11] T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill Science/Engineering/Math, Mar. 1997.

[12] P. Cheeseman, "In defense of probability," *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 1002—1009, 1985.

[13] A. Daneshkhah, "Uncertainty in probabilistic risk assessment: A review," University of Sheffield, Tech. Rep., Aug. 2004.

[14] N. Varachiu, "A fuzzy shapes characterization for robotics," in *Proceedings of the 6th International Conference on Computational Intelligence, Theory and Applications: Fuzzy Days*. Springer-Verlag, 1999, pp. 253–258.

[15] D. L. Shrestha and D. P. Solomatine, "Machine learning approaches for estimation of prediction interval for the model output," *Neural Networks*, vol. 19, no. 2, pp. 225 – 235, 2006, earth Sciences and Environmental Applications of Computational Intelligence.

[16] "Uncertainty quantification," 2009. [Online]. Available: http://en.wikipedia.org/wiki/Uncertainty_quantification

[17] V. Balasubramanian, S. Chakraborty, and S. Panchanathan, "Confidence estimation in pattern classification: An analysis with head pose estimation," School of Computing and Informatics, Arizona State University, Tech. Rep. TR-09-12, 2009.

[18] R. Krzysztofowicz, "The case for probabilistic forecasting in hydrology," *Journal of Hydrology*, vol. 249, no. 1-4, pp. 2–9, Aug. 2001.

[19] R. D. Lee, "Probabilistic approaches to population forecasting," *Population and Development Review*, vol. 24, pp. 156–190, 1998.

[20] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*, 2nd ed. Wiley-Interscience, Dec. 2007.

[21] A. Doucet, N. de Freitas, N. Gordon, and A. Smith, *Sequential Monte Carlo Methods in Practice*, 1st ed. Springer, Jun. 2001.

[22] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, Jun. 1965.

[23] G. J. Klir and T. A. Folger, *Fuzzy Sets, Uncertainty and Information*. Prentice Hall, 1988.

[24] G. Shafer, *Mathematical Theory of Evidence*. Princeton Univ Pr, Mar. 1976.

[25] F. Smarandache and J. Dezert, *Advances and Applications of DSmT for Information Fusion (Collected works), second volume*. Am. Res. Press, Sep. 2006.

[26] D. Dubois and H. Prade, *Possibility Theory*, 1st ed. Springer, Feb. 1988.

[27] E. Shortliffe, *Computer Based Medical Consultations: MYCIN*. American Elsevier, 1976.

[28] F. Lefvre, "Confidence measures based on the k-nn probability estimator," in *Sixth International Conference on Spoken Language Processing (ICSLP 2000)*, 2000.

[29] S. Eickeler, M. Jabs, and G. Rigoll, "Comparison of confidence measures for face recognition," in *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*. Washington, DC, USA: IEEE Computer Society, 2000, p. 257.

[30] G. Williams and S. Renals, "Confidence measures from local posterior probability estimates," *Computer Speech and Language*, vol. 13, pp. 395–411, 1999.

[31] J. Pinto and R.N.V.Sitaram, "Confidence Measures in Speech Recognition based on Probability Distribution of Likelihoods," 2005, hP Labs Technical Report HPL-2005-144.

[32] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, S. G, C. Goutte, A. Kulesza, N. Ueng, A. Sanchis, and N. Ueffing, "Confidence estimation for machine translation," in *In M. Rollins (Ed.), Mental imagery*. Yale University Press, 2004.

[33] B. Jiang, X. Zhang, and T. Cai, "Estimating the confidence interval for prediction errors of support vector machine classifiers," *J. Mach. Learn. Res.*, vol. 9, pp. 521–540, 2008.

[34] M. Marchand and J. S. Taylor, "The set covering machine," *J. Mach. Learn. Res.*, vol. 3, pp. 723–746, 2003.

[35] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," in *Machine Learning*. Morgan Kaufmann, 1998, pp. 260–268.

[36] P. M. Lee, *Bayesian Statistics: An Introduction*, 3rd ed. Wiley, 2009.

[37] M. Kuss, F. Jakel, and F. A. Wichmann, "Bayesian inference for psychometric functions," *Journal of Vision*, vol. 5, no. 5, pp. 478–492, 2005.

[38] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[39] J. Quionero-Candela, C. Rasmussen, F. Sinz, O. Bousquet, and B. Schlkopf, "Evaluating predictive uncertainty challenge," in *Machine Learning Challenges*, 2006, pp. 1–27.

[40] T. Fawcett, "An introduction to ROC analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[41] K. Proedrou, "Rigorous measures of confidence for pattern recognition and regression," Ph.D. dissertation, Royal Holloway College, University of London, 2003, advisor-Alex Gammerman.

[42] R. M. Neal, *Bayesian Learning for Neural Networks*, 1st ed. Springer, Aug. 1996.

[43] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*, 1st ed. Cambridge University Press, Jun. 2002.

[44] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[45] D. Haussler, "Probably approximately correct learning," in *Proceedings of the Eighth National Conference on Artificial Intelligence*. AAAI Press, 1990, pp. 1101–1108.

[46] "Probably approximately correct learning: Wikipedia," http://en.wikipedia.org/wiki/Probably_approximately_correct_learning, last Accessed: Sep 17, 2009.

[47] A. Moore, "A tutorial on PAC learning," http://www.autonlab.org/tutorials/pac.html, last Accessed: Sep 17, 2009.

[48] H. P. Kostas, K. Proedrou, V. Vovk, A. Gammerman, and S. T. Ex, "Inductive confidence machines for regression," *In Tapio Elomaa, Heikki Mannila, and Hannu Toivonen, Editors, Proceedings of the Thirteenth European Conference on Machine Learning*, vol. 2430, pp. 345—356, 2002.

[49] H. Papadopoulos, V. Vovk, and A. Gammermam, "Conformal prediction with neural networks," *Tools with Artificial Intelligence, IEEE International Conference on*, vol. 2, pp. 388–395, 2007.

[50] K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman, "Transductive confidence machines for pattern recognition," in *13th European Conference on Machine Learning*. Springer-Verlag, 2002, pp. 381–390.

[51] I. Nouretdinov, V. Vovk, M. V. Vyugin, and A. Gammerman, "Pattern recognition and density estimation under the general i.i.d. assumption," in *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*. Springer-Verlag, 2001, pp. 337–353.

[52] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, 1st ed. Cambridge University Press, Mar. 2000.

[53] W. Wiki, "Confidence interval," http://wapedia.mobi/en/Confidence_intervals, last Accessed: Oct 6, 2009.

[54] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, pp. 371–421, 2008.

[55] M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications (2nd ed.)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1997.

[56] V. Vovk, "On-line confidence machines are well-calibrated," in *FOCS '02: 43rd Symposium on Foundations of Computer Science*. Washington, DC, USA: IEEE Computer Society, 2002, pp. 187–196.

[57] S.-S. Ho and H. Wechsler, "Transductive confidence machine for active learning," *IJCNN*, vol. 2, pp. 1435–1440 vol.2, July 2003.

[58] V. Balasubramanian, S. Panchanathan, and S. Chakraborty, "Multiple cue integration in transductive confidence machines for head pose classification," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, 23-28 2008, pp. 1 –8.

[59] S.-S. Ho and H. Wechsler, "Query by transduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 9, pp. 1557–1571, 2008.

[60] A. Gammerman, I. Nouretdinov, B. Burford, A. Chervonenkis, V. Vovk, and Z. Luo, "Clinical mass spectrometry proteomic diagnosis by conformal predictors," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 2, p. Article13, 2008, PMID: 18673293.

[61] V. N. Balasubramanian, R. Gouripeddi, S. Panchanathan, J. Vermillion, A. Bhaskaran, and R. M. Siegel, "Support vector machine based conformal predictors for risk of complications following a coronary drug eluting stent procedure," in *Proceedings of the IEEE Conference on Computers in Cardiology*, 2009.

[62] G. Papadopoulos, P. Edwards, and A. Murray, "Confidence estimation methods for neural networks: a practical comparison," *Neural Networks, IEEE Transactions on*, vol. 12, no. 6, pp. 1287, 1278, 2002.

[63] H. Papadopoulos, "Inductive conformal prediction: Theory and application to neural networks," in *Tools in Artificial Intelligence*, August 2008, pp. 315–329.

[64] H. Papadopoulos, A. Gammerman, and V. Vovk, "Confidence predictions for the diagnosis of acute abdominal pain," in *Artificial Intelligence Applications and Innovations III*, 2009, pp. 175–184.

[65] F. Yang, H. zhen Wang, H. Mi, C. de Lin, and W. wen Cai, "Using random forest for reliable classification and cost-sensitive learning for medical diagnosis," *BMC Bioinformatics*, vol. 10, no. Suppl 1, p. S22, 2009.

[66] I. Nouretdinov, T. Melluish, and V. Vovk, "Ridge regression confidence machine," *In Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 385—392, 2001.

[67] T. Bellotti, Z. Luo, and A. Gammerman, "Strangeness minimisation feature selection with confidence machines," in *Intelligent Data Engineering and Automated Learning IDEAL 2006*, 2006, pp. 978–985.

[68] D. Hardoon, Z. Hussain, and J. Shawe-Taylor, "A nonconformity approach to model selection for svms," University College London, Tech. Rep., 2009.

[69] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, Sep. 1998.

[70] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *J. Mach. Learn. Res.*, vol. 9, pp. 371–421, 2008.

[71] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[72] V. Vovk, I. Nouretdinov, and A. Gammerman, "Testing exchangeability online," in *Proceedings of the Twentieth International Conference on Machine Learning*. AAAI Press, 2003.

[73] M. E. Matheny, F. S. Resnic, N. Arora, and L. Ohno-Machado, "Effects of SVM parameter optimization on discrimination and calibration for post-procedural PCI mortality," *J. of Biomedical Informatics*, vol. 40, no. 6, pp. 688–697, 2007.

[74] S. Vinterbo and L. Ohno-Machado, "A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction." *AMIA Symposium*, pp. 984–988, 1999.

[75] R. F. Harrison and R. L. Kennedy, "Artificial neural network models for prediction of acute coronary syndromes using clinical data from the time of presentation," *Annals of Emergency Medicine*, vol. 46, no. 5, pp. 431–439, Nov. 2005.

[76] L. Ohno-Machado and M. A. Musen, "Sequential versus standard neural networks for pattern recognition: an example using the domain of coronary heart disease," *Computers in Biology and Medicine*, vol. 27, no. 4, pp. 267–281, Jul. 1997.

[77] C. Stettler, S. Wandel, S. Allemann, and et al, "Outcomes associated with drug-eluting and bare-metal stents: a collaborative network meta-analysis," *The Lancet*, vol. 370, no. 9591, pp. 937–948, Sep. 2007.

[78] R. Gouripeddi, V. Balasubramanian, J. Harris, A. Bhaskaran, R. Siegel, and S. Panachanathan, "Predicting risk of complications following a drug eluting stent procedure: a svm approach for imbalanced data," *22nd IEEE International Symposium on Computer-Based Medical Systems*, pp. 984–988, Aug 2009.

[79] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "Svm and kernel methods matlab toolbox," Perception Systmes et Information, INSA de Rouen, Rouen, France, 2005.

[80] L. M. Brown and Y.-L. Tian, "Comparative study of coarse head pose estimation," in *IEEE Workshop on Motion and Video Computing*, Orlando, Florida, USA, 2002, pp. 125–130.

[81] S. Krishna, D. Colbry, J. Black, V. Balasubramanian, and S. Panchanathan, "A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired," in *ECCV Workshop on Computer Vision Applications for the Visually Impaired (CVAVI)*, 2008.

[82] V. Balasubramanian, S. Chakraborty, S. Krishna, and S. Panchanathan, "Human-centered machine learning in a social interaction assistant for individuals with visual impairments," in *Symposium on Assistive Machine Learning for People with Disabilities at Neural Information Processing Systems (NIPS)*, 2009.

[83] S. Krishna, V. Balasubramanian, and S. Panchanathan, "Enriching social situational awareness in remote interactions: Insights and inspirations from disability focused research," in *ACM Multimedia 2010 (Brave New Ideas)*, 2010.

[84] G. Little, S. Krishna, J. Black, and S. Panchanathan, "A methodology for evaluating robustness of face recognition algorithms with respect to variations in pose and illumination angle," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, USA, 2005, pp. 89–92.

[85] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 607–626, 2009.

[86] K. Bailly and M. Milgram, "2009 special issue: Boosting feature selection for neural network based regression," *Neural Netw.*, vol. 22, no. 5-6, pp. 748–756, 2009.

[87] S. Krishna, T. McDaniel, V. Balasubramanian, D. Colbry, and S. Panchanathan, "Haptic belt for delivering nonverbal cues to people who are blind/visually impaired," 2009.

[88] A. Ross and A. Jain, "Multimodal biometrics: An overview," in *Proceedings of the 12th European Signal Processing Conference (EUSIPCO)*, 2004.

[89] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. VDM Verlag, Jun. 2008.

[90] S. Marcel, C. McCool, S. Chakraborty, V. Balasubramanian, S. Panchanathan, J. Nolazco, L. Garcia, R. Aceves, and et al., "Mobile biometry (mobio) face and speaker verification evaluation," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010)*, 2010.

[91] V. Balasubramanian, S. Chakraborty, and S. Panchanathan, "Generalized query by transduction for online active learning," in *Proceedings of the International Conference on Computer Vision (ICCV 2009) Workshop on Online Learning for Computer Vision*, 2009.

[92] H. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen, "Multi-modal person identification in a smart environment," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–8.

[93] J. Nolazco-Flores and P. Garcia-Perera, "Enhancing acoustic models for robust speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[94] D. L. Renfrew, E. A. Franken, K. S. Berbaum, F. H. Weigelt, and M. M. Abu-Yousef, "Error in radiology: classification and lessons in 182 cases presented at a problem case conference," *Radiology*, vol. 183, no. 1, pp. 145–150, Apr. 1992, PMID: 1549661.

[95] S. Samuel, H. L. Kundel, C. F. Nodine, and L. C. Toto, "Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs," *Radiology*, vol. 194, no. 3, pp. 895–902, Mar. 1995, PMID: 7862998.

[96] K. Berbaum, E. A. Franken, R. T. Caldwell, and K. M. Schartz, "Can a checklist reduce SOS errors in chest radiography?" *Academic Radiology*, vol. 13, no. 3, pp. 296–304, Mar. 2006, PMID: 16488841.

[97] K. S. Berbaum, R. T. Caldwell, K. M. Schartz, B. H. Thompson, and E. F. Jr, "Does Computer-Aided diagnosis for lung tumors change satisfaction of search in chest radiography?" *Academic Radiology*, vol. 14, no. 9, pp. 1069–1076, Sep. 2007.

[98] M. Alzubaidi, J. Black, A. Patel, and S. Panchanathan, "Conscious vs subconscious perception as a function of radiological expertise," in *22nd International Symposium on Computer-based Medical Systems (CBMS)*, 2009.

[99] D. Manning, S. Ethell, T. Donovan, and T. Crawford, "How do radiologists do it? the influence of experience and training on searching for chest nodules," *Radiography*, vol. 12, no. 2, pp. 134–142, May 2006.

[100] L. Dempere-Marco, X. Hu, S. L. S. MacDonald, S. M. Ellis, D. M. Hansell, and G. Yang, "The use of visual search for knowledge gathering in image decision support," *IEEE Transactions on Medical Imaging*, vol. 21, no. 7, pp. 741–754, Jul. 2002, PMID: 12374312.

[101] X. Hu, L. Dempere-Marco, and G. Yang, "Hot spot detection based on feature space representation of visual search," *IEEE Transactions on Medical Imaging*, vol. 22, no. 9, pp. 1152–1162, Sep. 2003, PMID: 12956270.

[102] M. Antonelli and G. Yang, "Lung nodule detection using Eye-Tracking," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 2, 2007, pp. II – 457–II – 460.

[103] M. Kukar, "Transductive reliability estimation for medical diagnosis," *Artificial Intelligence in Medicine*, vol. 29, 2002.

[104] M. Kukar, "Quality assessment of individual classifications in machine learning and data mining," *Knowledge Information Systems*, vol. 9, no. 3, pp. 364–384, 2006.

[105] Y. Freund and R. E. Schapire, "A Decision-Theoretic generalization of on-Line learning and an application to boosting," 1995.

[106] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, Oct. 2000.

[107] P. Phillips, H. Moon, P. Rauss, and S. Rizvi, "The FERET evaluation methodology for face-recognition algorithms," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997, pp. 137–143.

[108] W. Kienzle, G. Bakir, M. Franz, and B. Scholkopf, "Face detection - efficient and rank deficient," *Advances in Neural Information Processing Systems*, 2005.

[109] K. W. Baran, J. M. Lasala, D. A. Cox, A. Song, M. C. Deshpande, M. V. Jacoski, and S. R. Mascioli, "A clinical risk score for prediction of stent thrombosis," *The American Journal of Cardiology*, vol. 102, no. 5, pp. 541–545, Sep. 2008.

[110] E. Romagnoli, F. Burzotta, C. Trani, M. Siviglia, G. G. L. Biondi-Zoccai, G. Niccoli, A. M. Leone, I. Porto, M. A. Mazzari, R. Mongiardo, A. G. Rebuzzi, G. Schiavoni, and F. Crea, "EuroSCORE as predictor of in-hospital

mortality after percutaneous coronary intervention," *Heart (British Cardiac Society)*, vol. 95, no. 1, pp. 43–48, 2009, PMID: 18208829.

[111] M. Singh, R. J. Lennon, D. R. Holmes, M. R. Bell, and C. S. Rihal, "Correlates of procedural complications and a simple integer risk score for percutaneous coronary intervention," *Journal of the American College of Cardiology*, vol. 40, no. 3, pp. 387–393, Aug. 2002.

[112] R. E. Shaw, H. Anderson, R. G. Brindis, R. J. Krone, L. W. Klein, C. R. McKay, P. C. Block, L. J. Shaw, K. Hewitt, and W. S. Weintraub, "Development of a risk adjustment mortality model using the american college of CardiologyNational cardiovascular data registry (ACCNCDR) experience: 19982000," *Journal of the American College of Cardiology*, vol. 39, no. 7, pp. 1104–1112, Apr. 2002.

[113] F. S. Resnic, L. Ohno-Machado, A. Selwyn, D. I. Simon, and J. J. Popma, "Simplified risk score models accurately predict the risk of major in-hospital complications following percutaneous coronary intervention," *The American Journal of Cardiology*, vol. 88, no. 1, pp. 5–9, Jul. 2001, PMID: 11423050.

[114] M. Moscucci, E. Kline-Rogers, D. Share, M. O'Donnell, A. Maxwell-Eward, W. L. Meengs, P. Kraft, A. C. DeFranco, J. L. Chambers, K. Patel, J. G. McGinnity, and K. A. Eagle, "Simple bedside additive tool for prediction of In-Hospital mortality after percutaneous coronary interventions," *Circulation*, vol. 104, no. 3, pp. 263–268, Jul. 2001.

[115] G. T. O'Connor, D. J. Malenka, H. Quinton, J. F. Robb, M. A. Kellett, S. Shubrooks, W. A. Bradley, M. J. Hearne, M. W. Watkins, D. E. Wennberg, B. Hettleman, D. J. O'Rourke, P. D. McGrath, T. Ryan, P. VerLee, and for the Northern New England Cardiovascular Disease Study Group, "Multivariate prediction of in-hospital mortality after percutaneous coronary interventions in 1994-1996," *J Am Coll Cardiol*, vol. 34, no. 3, pp. 681–691, Sep. 1999.

[116] S. G. Ellis, W. Weintraub, D. Holmes, R. Shaw, P. C. Block, and S. B. King, "Relation of operator volume and experience to procedural outcome of percutaneous coronary revascularization at hospitals with high interventional volumes," *Circulation*, vol. 95, no. 11, pp. 2479–2484, Jun. 1997.

[117] E. L. Hannan, M. Racz, T. J. Ryan, B. D. McCallister, L. W. Johnson, D. T. Arani, A. D. Guerci, J. Sosa, and E. J. Topol, "Coronary angioplasty volume-

outcome relationships for hospitals and cardiologists," *JAMA: The Journal of the American Medical Association*, vol. 277, no. 11, pp. 892–898, Mar. 1997.

[118] E. L. Hannan, D. T. Arani, L. W. Johnson, H. G. Kemp, and G. Lukacik, "Percutaneous transluminal coronary angioplasty in new york state. risk factors and outcomes," *JAMA: The Journal of the American Medical Association*, vol. 268, no. 21, pp. 3092–3097, Dec. 1992.

[119] Z. Xu, R. Jin, I. King, and M. Lyu, "An extended level method for efficient multiple kernel learning," in *Neural Information Processing Systems (NIPS)*, 2009.

[120] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[121] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, pp. 27–72, 2004.

[122] J. Ye, S. Ji, and J. Chen, "Multi-class discriminant kernel learning via convex programming," *J. Mach. Learn. Res.*, vol. 9, pp. 719–758, 2008.

[123] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, Oct. 2007.

[124] T. Xiong and V. Cherkassky, "A combined SVM and LDA approach for classification," in *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, vol. 3, 2005, pp. 1455–1459 vol. 3.

[125] B. Schlkopf and A. J. Smola, *Learning with kernels*. MIT Press, 2002.

[126] N. Cristianini, J. Shawe-taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 367–373.

[127] M. Gonen and E. Alpaydin, "Multiple kernel learning algorithms," Bogazii University, Bebek, Istanbul, Turkey, Tech. Rep., 2009.

[128] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *ICML '04: International Conference on Machine Learning*. New York, NY, USA: ACM, 2004, p. 6.

[129] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learn. Res.*, vol. 7, pp. 1531–1565, 2006.

[130] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, "New variants of bundle methods," *Math. Program.*, vol. 69, no. 1, pp. 111–147, 1995.

[131] R. Gouripeddi, V. Balasubramanian, S. Panchanathan, J. Harris, A. Bhaskaran, and R. Siegel, "Predicting risk of complications following a drug eluting stent procedure: A SVM approach for imbalanced data," in *Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on*, 2009, pp. 1–7.

[132] S. Mika, "Kernel fisher discriminants," Ph.D. dissertation, GMD FIRST, Berlin, Germany, 2002.

[133] R. Gouripeddi, V. Balasubramanian, S. Panchanathan, J. Harris, A. Bhaskaran, and R. Siegel, "Ranking predictors of complications following a drug eluting stent procedure using support vector machines," in *IEEE Computers in Cardiology*, 2009.

[134] R. Gouripeddi, "Kernel based approaches to provide decision support in interventional cardiology: Predicting complications following a drug eluting stent procedure," Master's thesis, Arizona State University, USA, 2010.

[135] Y. Fu and T. S. Huang, "Graph embedded analysis for head pose estimation," in *7th International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, 2006.

[136] B. Raytchev, I. Yoda, and K. Sakaue, "Head pose estimation by nonlinear manifold learning," in *17th International Conference on Pattern Recognition (ICPR04)*, Cambridge, UK, 2004.

[137] M. T. Wenzel and W. H. Schiffmann, "Head pose estimation of partially occluded faces," in *Second Canadian Conference on Computer and Robot Vision (CRV05)*, Victoria, Canada, 2005, pp. 353–360.

[138] V. N. Balasubramanian, J. Ye, and S. Panchanathan, "Biased manifold embedding: A framework for person-independent head pose estimation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR07)*, Minneapolis, USA, June 2007.

[139] V. N. Balasubramanian, S. Krishna, and S. Panchanathan, "Person-independent head pose estimation using biased manifold embedding," *EURASIP J. Adv. Signal Process*, vol. 2008, pp. 1–15, 2008.

[140] J. Heinzmann and A. Zelinsky, "3d facial pose and gaze point estimation using a robust real-time tracking paradigm," in *International Workshop on Automatic Face and Gesture Recognition*, Nara, 1998, pp. 142–147.

[141] M. Xu and T. Akatsuka, "Detecting head pose from stereo image sequence for active face recognition," in *International Workshop on Automatic Face and Gesture Recognition*, Nara, 1998, pp. 82–87.

[142] K. N. Choi, P. L. Worthington, and E. R. Hancock, "Estimating facial pose using shape-from-shading," *Pattern Recognition Letters*, vol. 23, pp. 533–548, 2002.

[143] Y. Hu, L. Chen, Y. Zhou, and H. Zhang, "Estimating face pose by facial asymmetry and geometry," in *IEEE Conference on Automatic Face and Gesture Recognition (AFGR04)*, 2004, pp. 651–656.

[144] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[145] H. Rowley, S. Baluja, and T. Kanade, "Neural network based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[146] S. Gundimada and V. Asari, "An improved snow based classification technique for head pose estimation and face detection," in *Proceedings of 34th Applied Imagery and Pattern Recognition Workshop (AIPR05)*, Washington, DC, 2005.

[147] Y.Wei, L.Fradet, and T.Tan, "Head pose estimation using gabor eigenspace modeling," in *IEEE International Conference on Image Processing*, Rochester, New York, USA, 2002.

[148] P. Fitzpatrick, "Head pose estimation without manual initialization," AI Lab, MIT, Tech. Rep., 2000.

[149] B. Tordoff, W. W. Mayol, T. d. Campos, and D. Murray, "Head pose estimation for wearable robot control," in *British Machine Vision Conference*, 2002, pp. 807–816.

[150] S. Ba and J. dobez, "A probabilistic framework for joint head tracking and pose estimation," in *IEEE International Conference on Pattern Recognition (ICPR04)*, 2004, pp. 264–267.

[151] S. O. Ba and J.-M. Odobez, "Evaluation of multiple cue head pose estimation algorithms in natural environments," in *IEEE International Conference on Multimedia and Expo (ICME05)*, Amsterdam, The Netherlands, 2005, pp. 1330–1333.

[152] S. Li, Q. Fu, L. Gu, B. Scholkopf, Y. Cheng, and H. Zhang, "Kernel machine based learning for multi-view face detection and pose estimation," in *IEEE International Conference on Computer Vision (ICCV01)*, vol. 2, 2001, pp. 674–679.

[153] M. Bichsel and A. Pentland, "Automatic interpretation of human head movements," Vision and Modeling Group, MIT Media Laboratory, Tech. Rep. 186, 1993.

[154] S. J. McKenna and S. Gong, "Real-time face pose estimation," *Real-Time Imaging*, vol. 4, pp. 333–347, 1998.

[155] S. Srinivasan and K. L. Boyer, "Head pose estimation using view based eigenspaces," in *IEEE International Conference on Pattern Recognition*, Quebec, Canada, 2002, pp. 302– 305.

[156] L. Chen, L. Zhang, Y. Hu, M. Li, and H. Zhang, "Head pose estimation using fisher manifold learning," in *IEEE International Workshop on Analysis and Modeling of Face and Gestures (AMFG03)*, 2003, pp. 203–207.

[157] Y. Zhu and K. Fujimura, "Head pose estimation for driver monitoring," in *IEEE Intelligent Vehicles Symposium*, Parma, Italy, 2004, pp. 501–505.

[158] N. Hu, W. Huang, and S. Ranganath, "Head pose estimation by non-linear embedding and mapping," in *IEEE International Conference on Image Processing (ICIP05)*, Genova, 2005, pp. 342–345.

[159] E. Xing, A. Ng, M. Jordan, and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Advances in Neural Information Processing Systems 15*, vol. 15, 2002, pp. 505–512.

[160] L. v. d. Maaten, E.O.Postma, and H. v. d. Herik, "Dimensionality reduction: A comparative review," University Maastricht, Tech. Rep., 2007.

[161] T. Hastie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 6, pp. 607–616, 1996.

[162] L. Yang, "The connection between manifold learning and distance metric learning," Carnegie Mellon University, PA, USA, Tech. Rep., 2007.

[163] X. He and P. Niyogi, "Locality preserving projections," in *NIPS*, 2003.

[164] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, 2003.

[165] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[166] S. Roweis and L. Saul, "Non-linear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[167] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, 17-21 2005, pp. 1208 –1213.

[168] D. d. Ridder, O. Kouropteva, O. Okun, M. Pietikainen, and R. P. Duin, "Supervised locally linear embedding," in *International Conference on Artificial Neural Networks and Neural Information Processing*, 333-341, 2003.

[169] N. Vlassis, Y. Motomura, and B. Krose, "Supervised dimension reduction of intrinsically low-dimensional data," *Neural Computation*, vol. 14, pp. 191–215, 2002.

[170] C.-G. Li and J. Guo, "Supervised isomap with explicit mapping," in *First IEEE International Conference on Innovative Computing, Information and Control (ICICIC'06)*, Beijing, China, 2006.

[171] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, "Non-linear dimensionality reduction techniques for classification and visualization," in *International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, 2002, pp. 645–651.

[172] X. Geng, D.-C. Zhan, and Z.-h. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 35, no. 6, pp. 1098–1107, December 2005.

[173] Q. Zhao, D. Zhang, and H. Lu, "Supervised lle in ica space for facial expression recognition," in *International Conference on Neural Networks and Brain (ICNNB05)*, Beijing, China, 2005, pp. 1970–1975.

[174] Y. Bengio, J. F. Paiement, P. Vincent, and O. Delalleau, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," 2004.

[175] H. Choi and S. Choi, "Robust kernel isomap," *Pattern Recognition*, vol. 40, no. 3, pp. 853–862, 2007.

[176] M.-C. Yeh, I.-H. Lee, G. Wu, Y. Wu, and E. Y. Chang, "Manifold learning, a promised land or work in progress," in *IEEE International Conference on Multimedia and Expo (ICME 05)*, Amsterdam, The Netherlands, 2005.

[177] X. Ge, J. Yang, T. Zhang, H. Wang, and C. Du, "Three-dimensional face pose estimation based on novel non-linear discriminant representation," *Optical Engineering Letters (SPIE)*, vol. 45, no. 9, 2006.

[178] T. McDaniel, S. Krishna, V. Balasubramanian, D. Colbry, and S. Panchanathan, "Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind," Oct 2008, pp. 13 –18.

[179] S. Panchanathan, S. Krishna, J. Black, and V. Balasubramanian, "Human centered multimedia computing: A new paradigm for the design of assistive and rehabilitative environments," 2008, pp. 1–7.

[180] S. Panchanthan, N. C. Krishnan, S. Krishna, T. McDaniel, and V. N. Balasubramanian, "Enriched human-centered multimedia computing through inspirations from disabilities and deficit-centered computing solutions." Vancouver, British Columbia, Canada: ACM, 2008, pp. 35–42.

[181] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[182] A. Dempster, "A generalization of bayesian inference," in *Classic Works of the Dempster-Shafer Theory of Belief Functions*, 2008, pp. 73–104.

[183] D. Dubois and H. Prade, "Possibility theory and its applications: a retrospective and prospective view," in *Fuzzy Systems, 2003. FUZZ '03. The 12th IEEE International Conference on*, vol. 1, 2003, pp. 5–11 vol.1.

[184] S. Beiraghi, M. Ahmadi, M. S. Ahmed, and M. Shridhar, "Application of fuzzy integrals in fusion of classifiers for low error rate handwritten numerals recognition," *Pattern Recognition, International Conference on*, vol. 2, p. 2487, 2000.

[185] B. G. Buchanan, *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Jun. 1984.

[186] editors Florentin Smarandache; Jean Dezert, *Advances and Applications of DSmT for Information Fusion (Collected works), second volume*. Am. Res. Press, Sep. 2006.

[187] G. Shafer, "Perspectives on the theory and practice of belief functions," *International Journal of Approximate Reasoning*, vol. 4, no. 5-6, pp. 323–362, 1990.

[188] E. Michaelsen and K. Jaeger, "Evidence fusion using the GESTALT-system," 2008, pp. 1–7.

[189] G. L. Rogova and V. Nimier, "Reliability in information fusion: Literature survey," in *Proceedings of the Seventh International Conference on Information Fusion*, P. Svensson and J. Schubert, Eds., vol. II. Mountain View, CA: International Society of Information Fusion, Jun 2004, pp. 1158–1165.

[190] J. Pearl, "Reasoning with belief functions: an analysis of compatibility," *Int. J. Approx. Reasoning*, vol. 4, no. 5-6, pp. 363–389, 1990.

[191] B. V. Dasarathy, *Decision Fusion*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1994.

[192] H. Ali, M. Antenreiter, P. Auer, G. Csurka, T. de Campos, Z. Hussain, J. Laaksonen, R. Ortner, K. Pasupa, F. Perronnin, C. Saunders, J. Shawe-Taylor, and V. Viitaniemi, "Description, analysis and evaluation of confidence estimation procedures for sub-categorization," Xerox Research Center Europe, Tech. Rep. D6.2.1, 2009.

[193] R. Ranawana and V. Palade, "Multi-Classifier systems: Review and a roadmap for developers," *Int. J. Hybrid Intell. Syst.*, vol. 3, no. 1, pp. 35–61, 2006.

[194] L. Lam, "Classifier combinations: Implementations and theoretical issues," in *Proceedings of the First International Workshop on Multiple Classifier Systems*. Springer-Verlag, 2000, pp. 77–86.

[195] L. H. C. Tippett, *The methods of statistics*, 4th ed. Dover, 1963.

[196] R. A. Fisher, *Statistical Methods for Research Workers.*, 15th ed. Macmillan Pub Co, Jun. 1970.

[197] B. Wilkinson, "A statistical consideration in psychological research," *Psychological Bulletin*, vol. 48, no. 3, pp. 156–8, May 1951.

[198] T. Liptak, "On the combination of independent tests," vol. 3, 1958.

[199] H. O. Lancaster, "The combination of probabilities: An application of orthonormal functions," *Australian & New Zealand Journal of Statistics*, vol. 3, no. 1, pp. 20–33, 1961.

[200] E. Edgington, "An additive method for combining probability values from independent experiments," 1972.

[201] G. Mudholkar and E. George, "The logit method for combining probabilities," 1979.

[202] I. J. Goods, "On the weighted combination of significance tests," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 17, no. 2, pp. 264–265, 1955.

[203] F. Mosteller and R. Bush, *Selected quantitative techniques*, ser. Handbook of Social Psychology, 1954.

[204] F. Pesarin, *Multivariate Permutation Tests With Applications in Biostatistics*, 1st ed. Wiley, Jun. 2001.

[205] T. M. Loughin, "A systematic comparison of methods for combining p-values from independent tests," *Computational Statistics & Data Analysis*, vol. 47, no. 3, pp. 467–485, October 2004.

[206] K. Delac and M. Grgic, "A survey of biometric recognition methods," in *Electronics in Marine, 2004. Proceedings Elmar 2004. 46th International Symposium*, 2004, pp. 193, 184.

[207] A. K. Jain, R. Bolle, and S. Pankanti, *Biometrics: Personal Identification in Networked Society*, 1st ed. Springer, 1999.

[208] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 14, pp. 4–20, 2004.

[209] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*, 1st ed. Springer, Oct. 2007.

[210] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity authentication," 1999.

[211] B. Duc, E. S. Bign, J. Bign, G. Matre, and S. Fischer, "Fusion of audio and video information for multi modal person authentication," *Pattern Recognition Letters*, vol. 18, pp. 835–843, 1997.

[212] N. Poh and J. Korczak, "Hybrid biometric person authentication using face and voice features," in *In Proc. AVBPA*, 2001, pp. 348–353.

[213] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 955–966, 1995.

[214] P. Verlinde and G. Chollet, "Comparing decision fusion paradigms using k-nn based classifiers, decision trees and logistic regression in a multi-modal identity verification application," 1999.

[215] T. H. Eugene, E. Weinstein, R. Kabir, and A. Park, "Multi-modal face and speaker identification on a handheld device," in *in Proc. Wkshp. Multimodal User Authentication*, 2003, pp. 120–132.

[216] H. K. Ekenel, Q. Jin, M. Fischer, and R. Stiefelhagen, "Isl person identification systems in the clear 2007 evaluations," pp. 256–265, 2008.

[217] R. Hu and R. Damper, "Fusion of two classifiers for speaker identification: removing and not removing silence," in *Information Fusion, 2005 8th International Conference on*, vol. 1, 2005, p. 8 pp.

[218] S. Palanivel and B. Yegnanarayana, "Multimodal person authentication using speech, face and visual speech," *Comput. Vis. Image Underst.*, vol. 109, no. 1, pp. 44–55, 2008.

[219] N. Fox, R. Gross, J. Cohn, and R. Reilly, "Robust biometric person identification using automatic classifier fusion of speech, mouth, and face experts," *Multimedia, IEEE Transactions on*, vol. 9, no. 4, pp. 701–714, 2007.

[220] K. Ban, K. Kwak, H. Yoon, and Y. Chung, "Fusion technique for user identification using camera and microphone in the intelligent service robots," in

Consumer Electronics, 2007. ISCE 2007. IEEE International Symposium on, 2007, pp. 1–6.

[221] M. Carrasco, L. Pizarro, and D. Mery, "Bimodal biometric person identification system under perturbations," in *Advances in Image and Video Technology*, 2007, pp. 114–127.

[222] D. Bolme, J. Beveridge, and A. Howe, "Person identification using text and image data," in *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, 2007, pp. 1–6.

[223] J. Luque, R. Morros, A. Garde, J. Anguita, M. Farrus, D. Macho, F. Marqus, C. Martnez, V. Vilaplana, and J. Hernando, "Audio, video and multimodal person identification in a smart room," in *Multimodal Technologies for Perception of Humans*, 2007, pp. 258–269.

[224] S. Marcel, C. McCool, S. Chakraborty, V. Balasubramanian, S. Panchanathan, J. Nolazco, L. Garcia, R. Aceves, and et al., "On the results of the first mobile biometry (mobio) face and speaker verification evaluation," *Lecture Notes on Computer Science (ICPR 2010 Contest Series)*, 2010.

[225] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[226] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[227] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *IEEE International Conference on Computer Vision*, 2007.

[228] F. Stentiford, "An attention based similarity measure with application to content based information retrieval," in *Storage and Retrieval for Media Databases*, 2003, pp. 20–24.

[229] F. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression," in *Picture Coding Symposium*, 2001, pp. 25–27.

[230] L. Itti, "Models of bottom-up attention and saliency," *Neurobiology of attention*, vol. 582, 2005.

[231] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," *Advances in neural information processing systems*, vol. 17, pp. 481–488, 2005.

[232] V. Navalpakkam and L. Itti, "Optimal cue selection strategy," *Advances in neural information processing systems*, vol. 18, p. 987, 2006.

[233] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[234] S. Nataraju, V. Balasubramanian, and S. Panchanathan, "Learning attention based saliency in videos from human eye movements," in *IEEE Workshop on Motion and Video Computing (WMVC)*, 2009.

[235] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[236] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey vision conference*, vol. 15, 1988, p. 50.

[237] T. Kadir and M. Brady, "Scale saliency: A novel approach to salient feature and scale selection," in *Visual Information Engineering, 2003. VIE 2003. International Conference on*, 2003, pp. 25–28.

[238] D. Ruderman, "The statistics of natural images," *Network: computation in neural systems*, vol. 5, no. 4, pp. 517–548, 1994.

[239] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu, "On advances in statistical modeling of natural images," *Journal of mathematical imaging and vision*, vol. 18, no. 1, pp. 17–33, 2003.

[240] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR07). IEEE Computer Society*, 2007, pp. 1–8.

[241] Z. Wang and B. Li, "A Two-stage Approach to Saliency Detection in Images," in *IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 965–968.

[242] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[243] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2, pp. 107–123, 2005.

[244] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, p. 360.

[245] A. Oikonomopoulos, I. Patras, and M. Pantic, "Human action recognition with spatiotemporal salient points," *IEEE Transactions on Systems, Man and Cybernetics-Part B*, vol. 36, no. 3, pp. 710–719, 2006.

[246] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, 2005, pp. 65–72.

[247] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, vol. 1, 2005.

[248] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple," in *Proceedings of CVPR2001*, vol. 1, 2001.

[249] A. Duchowski, "A breadth-first survey of eye-tracking applications," *Behavior Research Methods Instruments and Computers*, vol. 34, no. 4, pp. 455–470, 2002.

[250] L. Granka, T. Joachims, and G. Gay, "Eye-tracking analysis of user behavior in WWW search," in *Proceedings of the 27th annual international ACM*

SIGIR conference on Research and development in information retrieval. ACM New York, NY, USA, 2004, pp. 478–479.

[251] J. Salojarvi, I. Kojo, J. Simola, and S. Kaski, "Can relevance be inferred from eye movements in information retrieval," in *Proceedings of WSOM*, vol. 3, 2003, pp. 261–266.

[252] O. Oyekoya and F. Stentiford, "An eye tracking interface for image search," in *ETRA '06: Proceedings of the 2006 symposium on Eye tracking research & applications*. New York, NY, USA: ACM, 2006, pp. 40–40.

[253] O. Oyekoya and F. Stentiford, "Perceptual image retrieval using eye movements," *Int. J. Comput. Math.*, vol. 84, no. 9, pp. 1379–1391, 2007.

[254] W. Kienzle, F. Wichmann, B. Scholkopf, and M. Franz, "A nonparametric approach to bottom-up visual saliency," *Advances in neural information processing systems*, vol. 19, p. 689, 2007.

[255] W. Kienzle, F. Wichmann, B. Scholkopf, and M. Franz, "Learning an interest operator from human eye movements," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 17, 2006, p. 22.

[256] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[257] W. Kienzle, B. Scholkopf, F. Wichmann, and M. Franz, "How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements," *Lecture Notes in Computer Science*, vol. 4713, p. 405, 2007.

[258] S. Nataraju, "Visual attention based saliency in videos," Master's thesis, Arizona State University, USA, 2010.

[259] M. Alzubaidi, V. Balasubramanian, J. Black, A. Patel, and S. Panchanathan, "What catches a radiologists eye? a comprehensive comparison of feature types for saliency prediction," in *Proceedings of the SPIE Medical Imaging Conference*, 2010.

[260] S. Krishna, V. Balasubramanian, J. Black, and S. Panchanathan, "Person-specific characteristic feature selection for face recognition," in *Biometrics: Theory, Methods, and Applications*, 2008.

[261] S. Nataraju, V. Balasubramanian, and S. Panchanathan, "An integrated approach to visual attention modeling for saliency detection in videos," in *Springer book on Machine Learning for Vision-based Motion Analysis*, 2010.

[262] D. J. C. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992.

[263] M. Dredze and K. Crammer, "Active learning with confidence," in *Proceedings of ACL-08: HLT, Short Papers*, Columbus, Ohio, June 2008, pp. 233–236.

[264] Y. Freund, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," in *Machine Learning*, 1997, pp. 133–168.

[265] D. Shen, J. Zhang, J. Su, G. Zhou, and C. Tan, "Multi-criteria-based active learning for named entity recognition," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain: Association for Computational Linguistics, 2004, p. 589.

[266] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[267] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," *J Mach Learn Res.*, vol. 5, pp. 255–291, 2004.

[268] R. Kothari and V. Jain, "Learning from labeled and unlabeled data using a minimal number of queries," *Neural Networks, IEEE Transactions on*, vol. 14, no. 6, pp. 1496–1505, 2003.

[269] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal Of Machine Learning Research*, vol. 2, pp. 999—1006, 2000.

[270] P. Melville, S. M. Yang, M. Saar-Tsechansky, and R. J. Mooney, "Active learning for probability estimation using jensen-shannon divergence," in *ECML*, 2005, pp. 268–279.

[271] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Query by committee made real," in *NIPS*, 2005.

[272] D. Sculley, "Online active learning methods for fast label-efficient spam filtering," August 2007.

[273] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "Learning probabilistic linear-threshold classifiers via selective sampling," *Lecture Notes in AI*, vol. 2777, 2003.

[274] C. Monteleoni and M. Kaariainen, "Practical online active learning for classification," in *IEEE CVPR 2007*, 2007, pp. 1–8.

[275] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in *Proceedings of the 23rd international conference on Machine learning*. Pittsburgh, Pennsylvania: ACM, 2006, pp. 1081–1088.

[276] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 2001, pp. 107–118.

[277] T. Osugi, D. Kun, and S. Scott, "Balancing exploration and exploitation: A new algorithm for active machine learning," in *Proceedings of the Fifth IEEE International Conference on Data Mining*. IEEE Computer Society, 2005, pp. 330–337.

[278] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proceedings of the 17th International Conference on Machine Learning*, 2000, pp. 839–846.

[279] P. Mitra, C. Murthy, and S. Pal, "A probabilistic active support vector learning algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 413–418, 2004.

[280] C. Campbell, N. Cristianini, and A. J. Smola, "Query learning with large margin classifiers," in *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2000, pp. 111–118.

[281] J. Cheng and K. Wang, "Active learning for image retrieval with Co-SVM," *Pattern Recognition*, vol. 40, no. 1, pp. 330–334, 2007.

[282] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal Of Artificial Intelligence Research*, vol. 4, pp. 129—145, 1996.

[283] N. Roy and A. Mccallum, "Toward optimal active learning through sampling estimation of error reduction," *In Proceedings Of 18th International Conference. On Machine Learning*, pp. 441—448, 2001.

[284] A. Holub, P. Perona, and M. Burl, "Entropy-based active learning for object recognition," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, 2008, pp. 1–8.

[285] D. J. C. MacKay, "Information-Based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 590—604.

[286] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning.*, vol. 15, no. 2, pp. 201–221, 1994.

[287] M. Li, "Confidence-Based active learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 28, no. 8, pp. 1251–1261, 2006.

[288] M. Tang, X. Luo, and S. Roukos, "Active learning for statistical natural language parsing," *In Proceedings Of ACL 2002*, pp. 120—127, 2002.

[289] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," *17th ACM SIGIR conference*, pp. 3—12, 1994.

[290] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," *Machine Learning*, vol. 28, pp. 133—168, 1997.

[291] R. Liere and P. Tadepalli, "Active learning with committees for text categorization," *In AAAI-97*, pp. 591—596, 1997.

[292] A. K. Mccallum, "Employing EM in pool-based active learning for text classification," *In Proceedings Of The 15th International Conference On Machine Learning*, pp. 350—358, 1998.

[293] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proceedings of the Fifteenth International Conference on Machine Learning*, 1998, pp. 1–9.

[294] I. Dagan, "Committee-based sample selection for probabilistic classifiers," *Journal of Artificial Intelligence Research*, pp. 335—360, 1999.

[295] I. Muslea, S. Minton, and C. A. Knoblock, "Selective sampling with redundant views," in *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press / The MIT Press, 2000, pp. 621–626.

[296] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," *Journal of Machine Learning Research*, vol. 5, pp. 255–291, 2004.

[297] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proceedings of the 18th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2001, pp. 19–26.

[298] K. Nigam, A. Mccallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, pp. 103—134, 1999.

[299] D. Pelleg and A. Moore, "Active learning for anomaly and rare-category detection," *In Advances In Neural Information Processing Systems 18*, vol. 16, pp. 1073—1080, 2004.

[300] A. Schein and L. Ungar, *A-Optimality for Active Learning of Logistic Regression Classifiers*, 2004.

[301] C. A. Thompson, M. E. Califf, and R. J. Mooney, "Active learning for natural language parsing and information extraction," in *Proceedings of the Six-*

teenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 1999, pp. 406–414.

[302] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 208–215.

[303] H. N. Tat, H. T. Nguyen, and A. Smeulders, "Active learning using pre-clustering," *In Proceedings Of The 21st International Conference On Machine Learning*, pp. 623—630, 2004.

[304] A. Kapoor, G. Hua, A. Akbarzadeh, and S. Baker, "Which faces to tag : Adding prior constraints into active learning," in *International Conference on Computer Vision*, 2009.

[305] G. Qi, X. Hua, Y. Rui, J. Tang, and H. Zhang, "Two-Dimensional active learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008*, 2008.

[306] K. Brinker, "Incorporating diversity in active learning with support vector machines," *In Proceedings of the 20th International Conference on Machine Learning*, pp. 59—66, 2003.

[307] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Large-scale text categorization by batch mode active learning," in *Proceedings of the 15th International Conference on World Wide Web*. ACM, 2006, pp. 633–642.

[308] S. Hoi, R. Jin, and M. Lyu, "Batch mode active learning with applications to text categorization and image retrieval," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, no. 1, 2009.

[309] S. Hoi, R. Jin, J. Zhu, and M. Lyu, "Semi-supervised svm batch mode active learning for image retrieval," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–7.

[310] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proceedings of the 23rd International Conference on Machine learning*. ACM, 2006, pp. 417–424.

[311] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," in *Neural Information Processing Systems*, 2008.

[312] S. Chakraborty, V. Balasubramanian, and S. Panchanathan, "Learning from summaries of videos: Applying batch mode active learning to face-based biometrics," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR10) Workshop on Biometrics*, 2010.

[313] S. Tong, "Active learning: theory and applications," Ph.D. dissertation, Stanford University, CA, USA, 2001.

[314] C. Diehl and G. Cauwenberghs, "SVM incremental learning, adaptation and optimization," in *IJCNN*, vol. 4, 2003, pp. 2685–2690 vol.4.

[315] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[316] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. VDM Verlag, Jun. 2008.

[317] S. Chakraborty, V. Balasubramanian, and S. Panchanathan, "Batch mode active learning for biometric recognition," B. V. K. V. Kumar, S. Prabhakar, and A. A. Ross, Eds., vol. 7667, no. 1. SPIE, 2010.

[318] A. Arnold, R. Nallapati, and W. Cohen, "A comparative study of methods for transductive transfer learning," Oct 2007, pp. 77 –82.

[319] R. Luis, L. Sucar, and E. Morales, "Transfer learning for bayesian networks," in *Advances in Artificial Intelligence IBERAMIA 2008*, ser. Lecture Notes in Computer Science, H. Geffner, R. Prada, I. Machado Alexandre, and N. David, Eds. Springer Berlin / Heidelberg, 2008, vol. 5290, pp. 93–102.

APPENDIX A

PROOF RELATED TO DISCREPANCY MEASURE IN GENERALIZED QUERY BY TRANSDUCTION

In Chapter 6, as part of the Generalized Query by Transduction approach, we defined a matrix C which contains the absolute value of the pairwise differences between all the p-values obtained from the Conformal Predictions framework:

$$C_{ij}(P) = \left|P_i - P_j\right| \tag{A.1}$$

We now prove that for any given set of p-values, this matrix C will always have exactly one positive eigenvalue, which we use as a measure of disagreement in this work. We prove this claim in this appendix.

Lemma 1: An N by N square matrix which has -2 in all its superdiagonal entries, positive constants in all entries of the last row and 0 in all the other positions, always has a positive determinant.

Proof: Consider the case when $N = 2$. The matrix M_2 can be written as:

$$M_2 = \begin{bmatrix} 0 & -2 \\ d_1 & d_2 \end{bmatrix}$$

where d_1 and d_2 are positive constants. It is trivial to verify that this matrix has a positive determinant. Let us also consider the case when $N = 3$. The matrix M_3 is now given as:

$$M_3 = \begin{bmatrix} 0 & -2 & 0 \\ 0 & 0 & -2 \\ d_1 & d_2 & d_3 \end{bmatrix}$$

Again, it is easy to verify that this matrix has a positive determinant. Let us now assume that the proposition holds for some $N = n$, that is, let us assume that the following matrix M_n has a positive determinant $\det(M_n)$:

$$M_n = \begin{bmatrix} 0 & -2 & 0 & 0 & \dots & 0 \\ 0 & 0 & -2 & 0 & \dots & 0 \\ \vdots & & & & & \\ d_1 & d_2 & d_3 & d_4 & \dots & d_n \end{bmatrix}$$

Now, consider the case when $N = n+1$. The matrix M_{n+1} is given by:

$$M_{n+1} = \begin{bmatrix} 0 & -2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & -2 & 0 & \dots & 0 & 0 \\ \vdots & & & & & & \\ d_1 & d_2 & d_3 & d_4 & \dots & d_n & d_{n+1} \end{bmatrix}$$

The determinant of M_{n+1} is computed as:

$$det(M_{n+1}) = -(-2) \begin{vmatrix} 0 & -2 & 0 & \dots & 0 & 0 \\ 0 & 0 & -2 & \dots & 0 & 0 \\ \vdots & & & & & \\ d_1 & d_3 & d_4 & \dots & d_n & d_{n+1} \end{vmatrix}$$

The $n \times n$ matrix on the right is of a similar form as M_n, and hence its determinant, say $det(\hat{M}_n)$ is greater than zero. Therefore:

$$det(M_{n+1}) = 2 \times det(\hat{M}_n) > 0$$

since the d_is are arbitrary constants. Thus, we see that if the proposition holds for $N = n$, then it also holds for $N = n+1$. Therefore, by the principle of mathematical induction, we conclude that the proposition holds for all N. This proves Lemma 1.

Lemma 2: An N by N square matrix M where

- $M_{NN} = 0$

- $M_{ij} = -2$ for all i, j with $i = j$ except $i = N$ and $j = N$
- $M_{iN} = 1$, except when $i = N$
- M_{Nj} = a positive constant, except when $j = N$
- 0s in all other positions

has a positive determinant if N is odd and a negative determinant if N is even.

Proof: Let $N = 2$. The matrix M_2 is given by:

$$M_2 = \begin{bmatrix} -2 & 1 \\ d_1 & 0 \end{bmatrix}$$

Trivially, the determinant of M_2 is negative for positive d_1. Now, consider the case when $N = 3$. The matrix M_3 is given by:

$$M_3 = \begin{bmatrix} -2 & 0 & 1 \\ 0 & -2 & 1 \\ d_1 & d_2 & 0 \end{bmatrix}$$

It is easy to verify that the determinant of this matrix is positive for positive values of d_1 and d_2.

Let us assume that the proposition holds for $N = 2n-1$ and $N = 2n$, where n is a positive integer. Let us consider the matrix M_{2n+1}

$$M_{2n+1} = \begin{bmatrix} -2 & 0 & 0 & \dots & 0 & 1 \\ 0 & -2 & 0 & \dots & 0 & 1 \\ \vdots & & & & & \\ d_1 & d_2 & d_3 & \dots & d_{2n} & 0 \end{bmatrix}$$

The determininant is given by

$$det(M_{2n+1}) = (-2)\begin{vmatrix} -2 & 0 & 0 & \dots & 0 & 1 \\ 0 & -2 & 0 & \dots & 0 & 1 \\ \vdots & & & & & \\ d_2 & d_3 & d_4 & \dots & d_{2n} & 0 \end{vmatrix}$$

$$+1 \times \begin{vmatrix} 0 & -2 & 0 & \dots & 0 \\ 0 & 0 & -2 & \dots & 0 \\ \vdots & & & & \\ d_1 & d_2 & d_3 & \dots & d_{2n} \end{vmatrix}$$

The positive sign appears in front of 1 as it is in an odd position $2n+1$. The first determinant evaluates to a negative value as, by our assumption, the proposition holds for $N = 2n$ and the second determinant is positive by Lemma 1. Thus, $det(M_{2n+1})$ is positive.

Now, consider the matrix M_{2n+2}:

$$M_{2n+2} = \begin{bmatrix} -2 & 0 & 0 & \dots & 0 & 1 \\ 0 & -2 & 0 & \dots & 0 & 1 \\ \vdots & & & & & \\ d_1 & d_2 & d_3 & \dots & d_{2n+1} & 0 \end{bmatrix}$$

Its determinant is given as:

$$det(M_{2n+2}) = (-2)\begin{vmatrix} -2 & 0 & 0 & \dots & 0 & 1 \\ 0 & -2 & 0 & \dots & 0 & 1 \\ \vdots & & & & & \\ d_2 & d_3 & d_4 & \dots & d_{2n+1} & 0 \end{vmatrix}$$

$$-1 \times \begin{vmatrix} 0 & -2 & 0 & \dots & 0 \\ 0 & 0 & -2 & \dots & 0 \\ \vdots & & & & \\ d_1 & d_2 & d_3 & \dots & d_{2n+1} \end{vmatrix}$$

The negative sign appears in front of 1 as it is in an even position $2n+2$. The first determinant is positive since it is proved that the proposition holds for $N = 2n+1$ and the second determinant is positive by Lemma 1. Hence, $det(M_{2n+2})$ is negative. Thus, it is proved that if the proposition holds for $N = 2n-1$ and $N = 2n$, then it also holds for $N = 2n+1$ and $N = 2n+2$ and therefore, by the principle of mathematical induction, Lemma 2 holds for all N.

Lemma 3: For any given set of N p-values, the matrix C has a positive determinant if N is odd and a negative determinant if N is even.

Proof: Consider the case when $N = 3$ and let the three p-values be a, b and c. Let d_1 be the absolute difference between a and b and d_2 be the absolute difference between b and c. The matrix C_3 is given by:

$$C_3 = \begin{bmatrix} 0 & d_1 & d_1+d_2 \\ d_1 & 0 & d_2 \\ d_1+d_2 & d_2 & 0 \end{bmatrix}$$

Its determinant is given by:

$$det(C_3) = \begin{vmatrix} 0 & d_1 & d_1+d_2 \\ d_1 & 0 & d_2 \\ d_1+d_2 & d_2 & 0 \end{vmatrix}$$

Using the transformations Row1 = Row1 - Row2 and Row2 = Row2 - Row3, we have:

$$det(C_3) = \begin{vmatrix} -d_1 & d_1 & d_1 \\ -d_2 & -d_2 & d_2 \\ d_1+d_2 & d_2 & 0 \end{vmatrix}$$

$$= d_1 d_2 \begin{vmatrix} -1 & 1 & 1 \\ -1 & -1 & 1 \\ d_1+d_2 & d_2 & 0 \end{vmatrix}$$

Using the transformations Column1 = Column1 - Column2 and Column2 = Column2 - Column3, we have:

$$det(C_3) = d_1 d_2 \begin{vmatrix} -2 & 0 & 1 \\ 0 & -2 & 1 \\ d_1 & d_2 & 0 \end{vmatrix}$$

$$\Rightarrow det(C_3) > 0$$

by Lemma 2.

In general, let the N p-values be $a_1, a_2, a_3 \ldots a_N$. Let d_1 be the absolute difference between a_1 and a_2, d_2 be the absolute difference between a_2 and a_3 and so on. The determinant of the matrix C is then given by:

$$det(C) = \begin{vmatrix} 0 & d_1 & d_1+d_2 & \ldots & \sum d_i \\ d_1 & 0 & d_2 & \ldots & \sum d_i - d_1 \\ \vdots & & & & \\ \sum d_i & \sum d_i - d_1 & \sum d_i - (d_1+d_2) & \ldots & 0 \end{vmatrix}$$

Using the transformations Row1 = Row1-Row2, Row2 = Row2-Row3 ... Row(N-1) = Row(N-1)-RowN, we get:

$$det(C) = \begin{vmatrix} -d_1 & d_1 & d_1 & \dots & d_1 \\ -d_2 & -d_2 & d_2 & \dots & d_2 \\ \vdots & & & & \\ \sum d_i & \sum d_i - d_1 & \sum d_i - (d_1 + d_2) & \dots & 0 \end{vmatrix}$$

$$= d_1 d_2 \dots d_{N-1} \begin{vmatrix} -1 & 1 & 1 & \dots & 1 \\ -1 & -1 & 1 & \dots & 1 \\ \vdots & & & & \\ \sum d_i & \sum d_i - d_1 & \sum d_i - (d_1 + d_2) & \dots & 0 \end{vmatrix}$$

Using the transformations Column1 = Column1-Column2, Column2 = Column2-Column3 ... Column(N-1) = Column(N-1)-ColumnN, we get:

$$det(C) = d_1 d_2 \dots d_{N-1} \begin{vmatrix} -2 & 0 & 0 & \dots & 1 \\ 0 & -2 & 0 & \dots & 1 \\ \vdots & & & & \\ d_1 & d_2 & d_3 & \dots & 0 \end{vmatrix}$$

Hence, $det(C) > 0$ if N is odd and $det(C) < 0$ if N is even by Lemma 2. This proves Lemma 3.

Theorem: The matrix C, which contains the absolute values of the pairwise differences between all the p-values obtained from the Conformal Predictions framework, i.e. $C_{ij}(P) = \left|P_i - P_j\right|$, will always have exactly one positive eigenvalue.

Proof: Given an $n \times n$ matrix M, the characteristic polynomial of M is written as:

$$x^n - g_1 x^{n-1} + g_2 x^{n-2} - \dots + (-1)^n g_n = 0 \tag{A.2}$$

where the coefficient g_j is the sum of the determinants of all the sub-matrices of M taken j rows and columns at a time (symmetrically). Thus, g_1 is the trace of M (i.e.,

the sum of the diagonal elements), g_2 is the sum of the determinants of the $\frac{n(n-1)}{2}$ sub-matrices that can be formed from M by deleting all but two rows and columns (symmetrically), and so on. Continuing in this way, we can find g_3, g_4,... up to g_n, which of course is the determinant of the entire $n \times n$ matrix. Note that the n roots of the characteristic polynomial are the eigenvalues of the matrix M.

Now, let us assume that we have a similar characteristic polynomial for the given matrix C. From Descartes' rule of signs, if the terms of a single-variable polynomial with real coefficients are ordered by descending variable exponent, then the number of positive roots of the polynomial is either equal to the number of sign differences between consecutive nonzero coefficients, or less than it by a multiple of 2.

From Lemma 3, we know that $det(C) > 0$ if n is odd and $det(C) < 0$ if n is even. Hence, in the equation for the characteristic polynomial (Equation A.2), it is evident that g_1 is always positive (since it is the sum of sub-matrices of C, taking 1 row and column at a time, each of whose determinant is positive). Similarly, g_2 is always negative, g_3 is always positive, and so on. Substituting these signs in Equation A.2, we see that the characteristic polynomial for C has only one sign change between consecutive non-zero co-efficients (between the first and second terms). Thus, from Descartes' rule of signs, the matrix C always has only one positive eigenvalue (root of the characteristic polynomial). This proves the theorem.