Feature Space Augmentation: Improving Prediction Accuracy of Classical Problems in Cognitive Science and Computer Vison

Piyush Saxena
Marquette University

FEATURE SPACE AUGMENTATION: IMPROVING PREDICTION ACCURACY OF CLASSICAL PROBLEMS IN COGNITIVE SCIENCE AND COMPUTER VISON

by

Piyush Rai Saxena, B.S., M.S.

A Dissertation submitted to the Faculty of the Graduate School,
Marquette University, in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

Milwaukee, Wisconsin

December 2017

ABSTRACT

FEATURE SPACE AUGMENTATION: IMPROVING PREDICTION ACCURACY OF CLASSICAL PROBLEMS IN COGNITIVE SCIENCE AND COMPUTER VISION

Piyush Rai Saxena, B.S., M.S.

Marquette University, 2017

The prediction accuracy in many classical problems across multiple domains has seen a rise since computational tools such as multi-layer neural nets and complex machine learning algorithms have become widely accessible to the research community. In this research, we take a step back and examine the feature space in two problems from very different domains. We show that novel augmentation to the feature space yields higher performance.

Emotion Recognition in Adults from a Control Group: The objective is to quantify the emotional state of an individual at any time using data collected by wearable sensors. We define emotional state as a mixture of amusement, anger, disgust, fear, sadness, anxiety and neutral and their respective levels at any time. The generated model predicts an individual's dominant state and generates an emotional spectrum, 1x7 vector indicating levels of each emotional state and anxiety. We present an iterative learning framework that alters the feature space uniquely to an individual's emotion perception, and predicts the emotional state using the individual specific feature space.

Hybrid Feature Space for Image Classification: The objective is to improve the accuracy of existing image recognition by leveraging text features from the images. As humans, we perceive objects using colors, dimensions, geometry and any textual information we can gather. Current image recognition algorithms rely exclusively on the first 3 and do not use the textual information. This study develops and tests an approach that trains a classifier on a hybrid text based feature space that has comparable accuracy to the state of the art CNN's while being significantly inexpensive computationally. Moreover, when combined with CNN'S the approach yields a statistically significant boost in accuracy.

Both models are validated using cross validation and holdout validation, and are evaluated against the state of the art.

ACKNOWLEDGEMENTS

Piyush Rai Saxena, B.S., M.S.

I would like to thank my mother, my father, and my brother whose support along with God's grace enabled me to get to this point. I would like to thank my advisors and the entire dissertation committee for their guidance. I would like to thank Randy Kirk and Nithin Ramachandran from Direct Supply for their guidance, business acumen and hardware support without which is work could not be completed. I would like to thank the Graduate School and all the Marquette University administration.

TABLE OF CONTENTS

List of Figures

List of Tables

1.0 INTRODUCTION

The software design and engineering paradigm is at a tipping point. Industry centered static solutions are being replaced by human centered solutions. For instance, within a span of a decade we have moved from the static search engines based on lookup tables that were only changed during massive platform updates to today's dynamic search engines that make individual specific recommendations based on every bit of information from live click streams to social media. This is the beginning of data driven software design made possible by advances in cheap computational horsepower[1]-[5].

1.1 Human behavior and machine perception

Despite the advances we are far from the "prefect" prediction engines. This can be attributed to the quality and quantity of useful data. Human behavior is highly variable and volatile. The goal of artificial intelligence is to imitate human behavior. This leads us to the big question. How can a machine cope with the vast variability of human behavior? There are two simple solutions. The first one is, increasing the type of data that a machine learning algorithm trains on. Statistically, this would mean the addition of more predictors that explain the variability within the target classes. The second would be human validation of the machine learning outputs and modification of the feature space to represent the human validation. In both scenarios, we are augmenting the feature space. In the first scenario is an example of latitudinal augmentation (making a dataset wider by addition of additional predictors), while the second scenario is an example of longitudinal augmentation (making a table longer by increasing instances of recorded data)

Both techniques have their unique merits in increasing the prediction capability of artificial intelligence applications. Both try and mimic certain unique features of human perception. The first technique is an attempt to imitate the human ability of deductive reasoning. The more information we have as humans, higher is the fine-grained prediction accuracy. The second technique is an attempt to imitate the human ability of inductive reasoning, where the individuals experience and training are used to make predictions for a new scenario.

The premise can be better understood with a simple thought experiment. Imagine 2 graduate students are tasked with identifying the color of a bowling ball. To make things interesting, let the wavelength of the light coming from the ball (its color) be exactly half way between black and dark blue. Both graduate students from the Ubicomp Lab at Marquette University write wonderful computer vison applications to identify the color. The application developed by student 1 predicts black, and the application developed by the student 2 predicts dark blue. Which one is correct? The short answer is both are correct and incorrect at the same time. They are correct since they correctly imitate the creator's perception; they are incorrect since that is all they do. The model's accuracy in identifying color of the bowling ball ubiquitously is questionable at best. The tie is settled by a subject expert from imaging Physics who declares the ball to be black.

Now, let's look slightly beyond the mental warmup exercise. What will the consequences be if the students were tasked with identifying the emotions certain videos might induce? It is not possible to call a domain expert for whatever models our bright graduate students come up with. The variability of emotion perception is a great example of a problem where the prediction class is highly variable. A cliff diving video might induce

amusement in an adrenalin junkie but extreme fear in an individual with Acrophobia despite the physiological state of the individuals being nearly identical. Machine learning requires static prediction fields. Thus, the model would either predict “Amusement” or “Fear”, thus it will fail to predict the correct emotional response for either the adrenalin junkie or the Acrophobe.

1.2 Longitudinal Feature Space Augmentation

To solve this conundrum, we introduce the longitudinal feature space augmentation. To understand this, let’s take the help of another thought experiment. We were all children once upon a time. Imagine a child who has reached the age of reason, 7. How does that child act? Most actions are learnt from his experiences with his peer group which is indicative of the social norms prevalent (also a function of time). He perceives the kids consuming alcohol behind the school to be cool, since that is the perception amongst his peers. Parental guidance intervenes and now the child’s perception of alcohol is altered to being a bad substance. What would the child do if he saw the same kids now smoking cigarettes? The child does not associate it with being “cool” despite what the perception of his peers might be. He can now connect the alcohol intake to cigarette smoking. Both being unsuitable for his age. This perception is validated and strengthened by parental guidance. Over time, these validated perceptions shape his behavior as an adult.

1.3 Latitudinal Feature Space Augmentation

Now, let’s assume the same scenario occurs at a boarding school. The child might not have parental validation except for a goal to perform well in school. The child is aware of the possible target class (success vs. failure). The “alcohol consumption” and “cigarette

smoking" are now a part of the feature space that the child uses to classify his peers into the "successful" and the "failure" target classes. This is a great candidate for latitudinal feature space augmentation.

These principles can be applied to machine learning problems. In this dissertation, we will present solutions to two unique problems. While these problems and the presented solutions are novel works of research in their own rites, we show the value in feature space augmentation.

The first problem from the cognitive sciences domain is *real-time prediction of the human emotional state using physiological data from sensors*. Here we will motivate and evaluate the application of longitudinal data augmentation. The second problem is from the computer vison domain is *image classification of industrial equipment* where we create a novel hybrid feature space that employs latitudinal feature space augmentation to boost the prediction accuracy.

1.4 Dissertation Outline

This dissertation in divided into 7 chapters. Chapter 1, Introduction motivates the thinking behind feature space augmentation. Here we describe with simple examples the concepts of latitudinal feature space augmentation and longitudinal feature space augmentation and their connection to inductive and deductive reasoning in humans. We also present the motivation behind the work and the novel contributions of this dissertation.

Chapter 3 looks at Problem1, emotion modelling and recognition in a control group. Chapter 3 is further divided into 9 subsections that talk about related works and taxonomy, experiment design and data collection, research questions, survey data analysis, predicting

dominant emotional state in a 6 category classification problem (6 emotional states) and a 3 category prediction problem, predicting the emotional spectrum for a 7(6 dominant emotional states and Anxiety) and 4 output categories(6 dominant emotional states and Anxiety) and finally a proposed real time application. The feature space augmentation is a part of this sub-section.

Chapter 4 looks at problem 2, Image classification using an augmented feature space. This section is further divided into 5 subsections including related works and taxonomy, research objectives, methodology, algorithm pipeline and evaluation.

Chapter 5 includes a conclusion section that talks about our contributions and broader impact of this work.

Chapters 6,7 and 8 are the references, bibliography and the appendix respectively.

2.0 Motivation

Artificial intelligence is impacting our lives every day. Despite the concerns regarding the impact of AI on the job market there are certain application where the urgent need for AI and its potential game changing influence cannot be ignored. In this dissertation, we attempt to solve two such problems.

2.1 Predicting the emotional state of an Individual

Background: Ubicomp Lab at Marquette University partnered with the Milwaukee PEERS project in 2014 to understand the mathematics behind emotion perception in an ASD population [6], [7]. "*PEERS is an evidence-based, manualized, 14-week (16 weeks for young adults), outpatient treatment program developed at the University of California at Los Angeles. Dr. Van Hecke is certified by UCLA to provide the PEERS program at Marquette University."*[6] The subjects include teens (ages 11-16) and young adults (ages 18-28) with Autism spectrum disorder. Our lab collected physiological data and facial images from all PEERS sessions since 2014. The goal of the data collection was to understand and model emotion perception in ASD population. Specifically, event detection to recognize the occurrence of anxiety. The results from facial recognition did not achieve the accuracy required for clinical testing while the physiological data could not be used for event detection due to the lack of target classes. Thus, we designed an experiment from scratch which amongst other things allows for prediction of anxiety near real-time (with a lag of 60 seconds). Moreover, ASD is a spectrum disorder. There is variability in emotion perception amongst individuals. To account for this variation, we propose the longitudinal feature space augmentation based on human input. As a ground work for future research in mental

disorders we present and evaluate (against the state of art) a novel framework for individual specific emotion modelling. Moreover, there is merit in modelling emotional perception in general.

1. According to a recent study by the CDC and the National institute of health statistics the rate of Autism in the United States is 1 in 45[8]. This makes Autism one of the fastest growing developmental disorders.
2. The rate of Autism increased by over 119% between 2000-2010[9]
3. An economic forecasting study conducted at University of California Davis estimates the current cost (direct medical, direct non-medical and productivity) related to ASD to rise from 268 billion USD in 2015 to 461 billion USD in 2025[10]. This could account for about 4% of the United States GDP. If the rate of increase does not taper, the costs associated with ASD will exceed diabetes and ADHD by 2025.
4. Early intervention has significant cost benefits and benefits to the individual in leading a fulfilling life [11], [12].
    a. Over 65% of the cost associated with Autism is Adult spending[13].
    b. These costs can be reduced by 2/3 if an early intervention is provided[14]

2.2 Novel Image Classification using a Hybrid Feature Space

Direct Supply is an industry in senior living. The senior living industry has seen massive cuts in spending over the years while more of the demographic moves into assisted living and skilled nursing facilities. These facilities have industrial equipment (assets) that are managed by a service provider. Inventory management is a critical gear in this workflow. It is often tedious and requires significant time commitment from the facility manager. The business

need for an automatic inventory system was presented in Summer 2016. The objective was to train a computationally efficient (given limitations of mobile phones) classifier that could distinguish between the 15 asset categories with a high accuracy. Current state of the art for image classification include convolutional neural nets(CNN) that are highly compute intensive and reduce an image to a minimum of 1x1000 array. This makes retraining the network very expensive and limits mobile phone use.

2.3 Novel Contributions

Both works make novel contributions to the current body of scientific work in their respective domains.

2.3.1 Predicting the emotional state of an Individual

We make the following contributions in through this research,

i. The design and implementation of a system that can distinguish between dominant emotional states (using physiological data)
    a. 6-category classification (87.4% accuracy)
    b. 3-category classification (92% accuracy)

   The reported accuracy is the highest for the number of prediction classes among all surveyed works.

ii. The design and implementation of a system to predict the emotional spectrum (levels of all 6 dominant emotions and anxiety) for an individual. This is a completely novel work with nothing similar found in the literature review.

iii. A novel feature space augmentation algorithm that allows the feature space to be tailored to the emotion perception unique to every individual. This is a completely novel work with nothing similar found in the literature review.

iv. An in-depth study that spatially locates the emotions (based on feature space) and identifies variability in emotional perception and over laps between dominant emotions.

v. We achieve functional accuracy using PPG alone (the technology in most modern hears rate monitors). The accuracy is 87.4 % in a 6-category classification and 92% in a 3-category classification was achieved. This is significant since our system can be implemented using only a 25$ wearable sensor watch (heartrate only). This is significant cost savings when compared to other commercial systems that cost upwards of 1500$. Our technology can be massively scaled due to the low costs.

vi. We will share a data set of over 600 instances (each instance contains a 1x7 survey response and 5 physiological time series and a class), the raw data set with the videos used for the study and the data collection application as part of supplemental materials. This is the largest dataset (number of subjects) recorded to date.

2.3.2 Novel Image Classification using a Hybrid Feature Space

i. Algorithm to re-encode based on a text based feature space. This feature space has unique properties. It performs as well as the state of the art CNN's while training a classifier on a 15-dimension feature space compared to 1000's of dimensions in the CNN. This leads to significant computational efficiency (training times and prediction speeds)

ii. Higher Information hybrid feature space- the addition of the text based features leads to a statistically significant information gain creating a classifier that boosts the classification accuracy of the state of the art image classification algorithms including Neural nets and key point extractors.

3.0 PROBLEM 1

Problem Statement: Emotion modelling and prediction uses real time wearable sensors. We are surrounded by an IOT web where our interactions with the digital world, are used to predict our actions to some end. Most of these predictions are centered around the physical world, such as activity recognition and fall detection, in this study, we focus our attention on the psychological world and emotional state of the individual. During an experimental study with 85 participants we induced specific emotions using audio-visual stimulus and collected physiological data, including heart rate, blood volume pressure(BVP), inter beat interval(IBI) and electrodermal activity(EDA) along with a self-report indicating the levels of 6 emotional states, Amusement, Anger, Sad, Disgust, Fear and Neutral. Additionally, we recorded a self-reported score for Anxiety. The videos used to induce emotions were validated in a recently published study in Psychology. The data collected was used to create models that identify the dominant emotional state and predict the emotional spectrum (levels of all emotional states) of an individual. An iterative learning framework is implemented to account for variability in emotional perception (the same stimulus might induce opposite emotional responses in different individuals) and generate an emotional spectrum unique to the individual. We report over 87 percent classification accuracy in a 6-category classification (dominant emotional state) and over 92 percent accuracy in a 3-category (positive, negative, neutral) classification. The emotional spectrums for the 6-state classification were modelled using the self-report data and the physiological data recorded during the experiment. The model was implemented in a server based application to identify the dominant emotional state and produce the emotion spectrum using 60 second streams of physiological data collected using wearables. Finally, we outline key implications for the design and implementation of a real-time

application with an iterative learning module for the prediction of the dominant emotional state and the corresponding emotion spectrum, unique to an individual's emotion perception.

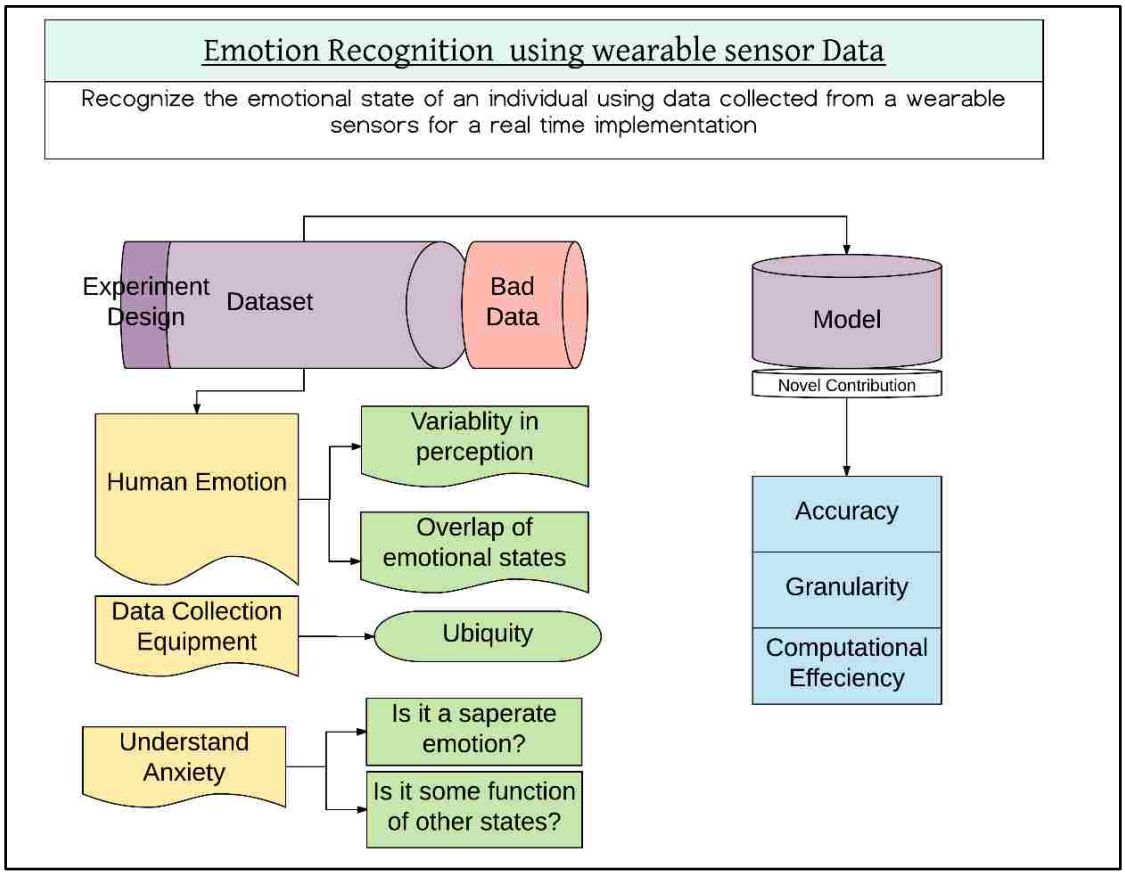


Figure 1: Research Objectives

The paradigm of situation-aware applications focuses at providing solutions unique to the situation of an individual. Significant effort has been put towards understanding the components that can define the state of an individual [15]-[20]. As a general example, two individuals that walk into the same room might have entirely different emotional states. One could be the boss and other the employee about to be fired. The major challenge from a situation awareness perspective is how we can identify those unique states. Moreover, we

need to establish if one dominant state can sufficiently define the emotional state. With respect to the boss-employee example, the employee would possibly experience a multitude of emotions such as anger, fear, anxiety and disgust. Which one of these states would be the dominant emotion would vary based on individual's experiences and perception. In this study, we demonstrate an approach to identify six dominant emotional states and the emotion spectrum unique to an individual (1x6 array with levels for each emotional state) using psychological data recorded during an experiment. Moreover, we propose and implement an iterative learning framework that allows for the general model to evolve and tailor itself to itself to the emotion perception unique to an individual. The current state of the art real-time emotion detection leverages the advances in computer vision to detect small changes in facial features [20]-[22]. This approach, while being highly efficient, is limited to the times when a facial image of acceptable resolution is available. Our approach allows for continuous emotion monitoring using a low-cost wearable heart rate monitor. This will allow us to tie the state of an individual to the data collected from the individual IOT touchpoints, thus creating a holistic picture that unites the digital world to human psychology.

3.1 Related Work and Taxonomy

We compare our system to the state of the art research in the public domain based on the following characteristics,

Prediction accuracy: Prediction accuracy refers to the percent of instances classified correctly by a machine learning algorithm. Our system achieves a maximum of **92 % accuracy for a 3-category** classification (Positive, Negative and Neutral) and an **87.4% accuracy for a 6-category** classification (Amusement, Sad, Fear, Neutral, Anger, Disgust).

This is referred to the valance level in a lot of research. The arousal level (low, medium, high) is a continuous variable (0-100) in our work.

Prediction granularity: Granularity refers to the number of predicted classes. Our work includes,

- 6 category Classification –valence level
- 3 category Classification - valance level
- 1x7 Emotion spectrum with scores for valence level categories and Anxiety

Computational efficiency: Computational efficiency can be derived from the time taken for feature extraction, training a classifier and prediction times. Since this information is not readily available for most of the published research, we will compare the dimension of the feature space as a measure for computational efficiency (, higher feature dimensions require higher training times, more complex models and higher prediction times). Our feature space is 3 dimensional.

Scope of Real time implementation: Real time implementation depends on multiple factors. The most important being a continuous data stream (use to predict state). This could be physiological data, facial images, audio data and data from social media. The second critical requirement is mobility. It is not practical to expect mass use of a system that requires multiple sensors strapped to an individual. Wearable sensor used in our work, E4 Empatica wristband provides a great balance between a continuous data stream and non-intrusive and non-obstructive data collection[23].

Implementation of an iterative learning framework: This is an element unique to our work alone. It allows the AI predicting emotional states to rapidly evolve and tailor the feature space uniquely to an individual's emotion perception. Moreover, an iterative learning framework acknowledges that emotions do not exist discretely; there is overlap between affective states (proven in the analysis section). Thus, at any time there exists an emotional spectrum, a 1x 7 vector with the proportion of each affective state and anxiety.

The autonomic nervous system regulates the unconscious actions of the body. It includes two primary divisions: Sympathetic nervous system (SNS) and Parasympathetic nervous system (PSNS). Sympathetic nervous system- like other divisions- operates through a string of tightly interconnected neurons[24], [25]. Albeit a significant portion is within the CNS (Central Nervous System), the Sympathetic nervous system is commonly considered as one of the components of the PNS (Peripheral Nervous System). The fundamental process of the sympathetic nervous system is to excite or stimulate the fight-or-flight response of the human body. On the other hand, the fundamental process of the parasympathetic nervous system which is to stimulate the "feed and breed" response, and after that, to the "rest-and-digest" response of the human body[26], [27]. From a computer scientist's perspective, we can think of the SNS and the PSNS as systems that counter each other. For instance, how angry one might become is governed by the SNS and the rate at which the individual calms down is governed by PSNS. Hence, it is theoretically possible to model one system if the response from the other is known.

Significant work has been done in the affective computing domain aimed at identifying affective states using data from wearable sensors, facial recognition, audio signals and even social media. A 2017 work by Ragot et al. compares the effectiveness of laboratory

sensor BIOPAC-MP150 to a wearable sensor Empatica E4 in terms of emotion recognition accuracy [23], [28], [29]. The study with 19 subjects validates the use of wearable sensors for emotion recognition outside the laboratory based on the physiological response recorded by both systems to the International Affective Picture System (IAPS) database [30]. The data was categorized under three levels of valence, positive, negative and neutral and three levels of arousal, high, medium and low. Nine specific features including HR, AVNN, SDNN, and rMSSD, Pnn50, LF, HF, RD and AVSCL were extracted to train the machine learning classifier. The authors used an 80-20 split with cross validation, reporting 66% accuracy for the valence level and 70% accuracy for the arousal level. Minhad et al. presented a study that uses physiological sensor data (specifically skin conductance) to model the emotional states of happiness, sadness, disgust, fear and anger [31]. The authors report an accuracy of over 70 % across the 5 categories. A 2016 study titled An Emotion Recognition System Based on Physiological Signals Obtained by Wearable Sensors by He et.al conducted experiments aimed at inducing joy, sadness, anger and pleasure on 11 subjects [32]. Electrocardiogram (ECG) and respiration (RSP) were recorded. The authors use a 145-dimension feature space to for classification with a SVM. The recognition accuracy was 81.82, 63.64, 54.55, and 30.00 % for joy; sadness, anger, and pleasure, respectively (average accuracy of 57.34%). Maaoui et al. published their work on emotion recognition in 2010[33]. The study used a linear discriminant classifier trained on a 30-dimensional feature space to predict 6 valence levels with an accuracy of 92%. The features space is derived from 50 second recordings of Blood volume pulse, Electromyography, Skin conductance, Skin Temperature and respiration for a subject pool of 10 participants. The features extracted are taken from Picard et al. [34] . While this work has a higher accuracy than our valence level

predictions (87.4%) the training set for this work includes 6 instances 50 second time series for each emotion compared to over 65 instances of 50 seconds-120 seconds time series for each emotion in our work. The 2001 work by Picard et al. at is one of the most iconic works that laid the foundation of emotion recognition using physiological data [34]. The proposed feature extraction is still widely used (along with other features) in the research community. The authors gathered data from 4 sensors measuring electromyogram, blood volume pressure, skin conductance and respiration. 6 features are extracted from each time series. The feature space was used to predict the emotional states including Neutral, Anger, Grief, Joy and Reverence using Fisher projection and Sequential floating forward search. The 5-category classification yielded a 46.3% accuracy and a 3-category classification (Anger, Joy and Reverence) yielded an 88.3 % accuracy.

In Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System, Haag et al. utilize EMG, EDA, ST, ECG, Respiration to create a feature space contain the running mean, running standard deviation and slope of the signals to predict the valance and the arousal level[30], [35]. A neural net is used as a predictor. Results were evaluated based on a tolerance of 10 and 20 percent (i.e. if the prediction was within the tolerance, the instance was classified correctly). A 10 percent bandwidth (tolerance) leads to 90% classification accuracy of valance levels and 63% accuracy in arousal levels. The major concern we have with this research is that the entire study is based on data collected from 1 individual. Moreover, valance and arousal describe a plane where all emotions lie and not the location (coordinates) of emotional states. Thus, the classification is abstract (high valence-high arousal, low valance-low arousal, high valence-low arousal and low valence-high arousal).

To recognize emotion in speech, J.P. Arias et al. presented a shape-based modeling of fundamental frequency contour in 2014 [36]. Here, with the help of the functional data analysis, they suggested neutral reference models identify emotions in the fundamental frequency and experienced considerably higher accuracy. This approach was applied to identify the most emotionally striking segments, and by using a natural database, verified at the sub-sentence level.

Zhang et al. delineated the process to detect emotions (happy, sad and neutral) by using the Kinect 3D Facial Points[37]. For this purpose, the authors used 1347 3D facial points by the Kinect V2.0, selected the key points, and performed the feature extraction. Machine learning classifiers were employed to create the emotion identification models.

Soleymani et al. presented a continual emotion detection approach using a unique combination of facial expressions and EEG signals[38]. In this approach, each subject was let to view a short emotional video. Then, multiple annotators were set to continually provide the valence levels by following the frontal facial videos of each subject. Here, besides the facial fiducial points, the authors used power spectral features from EEG signals as features to identify the valence levels for each of the frames. In [39], Claudio Loconsole et al. proposed a unique methodology to extract facial features and recognize the facial emotions automatically with high accuracy. Employing real-time face tracker, they extracted two distinct features such as linear features and eccentricity. Then, these features trained the machine learning classifiers. This method allowed 6 primary Ekman's emotions classification in real time without requiring any prior information of facial traits and manual intervention.

To detect human emotion, M. Liu et al. combined multiple kernel methods on the Riemannian Manifold at [40]. In this approach, each of the video clips was described by the

covariance matrix, linear subspace, and Gaussian distribution. These images set models were observed as points residing on Riemannian manifolds. After that, for similarity measurement, Riemannian kernels are applied on these models accordingly.

Veenendal et al. analyzed the emotion recognition in a group and crowd ambiance[41]. In this course, the edge detection was practiced with a Mesh Superimposition to extract the regarding features. The authors applied the feature movement (based on the shift from the reference point) to track across the strings of the images from a color channel. Furthermore, to validate their approach, they captured video of a group of subjects on spontaneous emotions while watching sports competitions.

R. Rakshit et al. proposed emotion detection using HRV (Heart Rate Variability) features obtained from the PPG (photoplethysmogram) signals in[42]. In this study, a Pulse Oximeter was used to collect heart rate signals and detect emotional states. The HRV features are obtained from both the time and frequency domain and then employed for emotion classification. The researchers extracted features from the PPG signal received in the baseline neutral and the emotion elicitation phase. Employing the HRV features, as well as the standard video stimuli, they analyzed three emotions: happy, neutral, and sad.

Rao et al. proposed an affective topic model for the social-emotion recognition regarding the social media platform and offered an intermediate layer to meet the objective [43]. This model can be implemented to classify (or incorporate) the social-emotions regarding the unlabeled documents (texts or records) aimed at developing a social-emotion lexicon.

Lei et al. concentrated on building a social-emotion identification approach for the online reports leading to social-emotion lexicon generation [44]. It also focused on emotion-ambiguity detection and the context-dependence of the sentiment orientations.

To enhance the multimedia Content, F. Yu et al. presented an experimental research on the speech-based emotion recognition in [45]. The primary dataset is a collection of written texts comprising of emotional speech with 721 short speeches. These speeches express four target emotions (happiness, anger, neutral, and sadness. The investigation revealed that the emotion prediction based on textual data alone is not accurate.

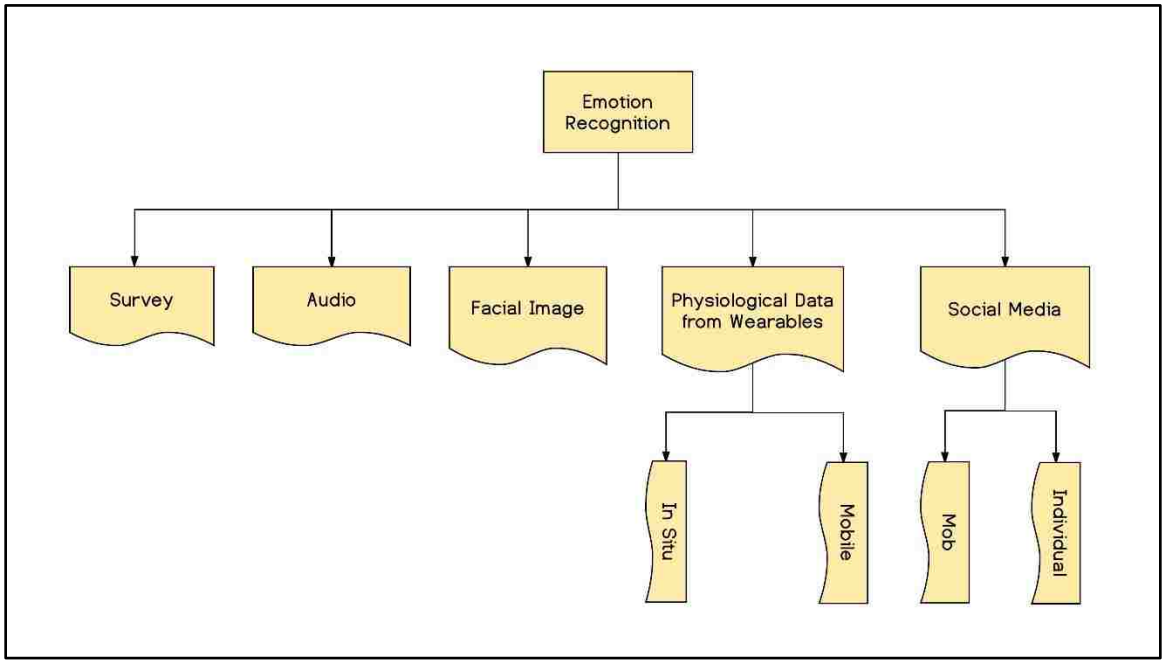


Figure 2: Taxonomy of emotion detection

Based on the literature survey we can break down the Emotion recognition application into 5 major categories listed in *Figure 2: Taxonomy of emotion detection*. While surveys, audio, facial

images and social media have their merits, they are not suitable for the problem we are trying to solve. Our goal is continuous emotion recognition, possible exclusively via mobile wearable sensors. The data flow from audio and social media is not continuous. While 24-hour video feeds of surveys (even at discrete time intervals) are not practical. Moreover, the social media data and video feeds (for facial image based emotion recognition) pose significant privacy and security risks for an individual. Moreover, a significant work done using wearable sensors involves devices such as electrocardiograph and respiration rate monitors which are not mobile. A wearable senor watch is a practical solution (if it can make accurate predictions) for a system that can be scales across a wide variety of populations.

3.2 Experiment Design and Data Collection

We conducted an experimental study to collected survey responses and physiological time series data in response to a data set of videos. 85 subjects between the ages of 18-24 were recruited for the study. The data set used was leveraged from Hewig et al. that recorded survey responses to classify the videos into dominant emotional states [46]. These states include Amusement, Anger, Neutral, Sad, Fear and Disgust.

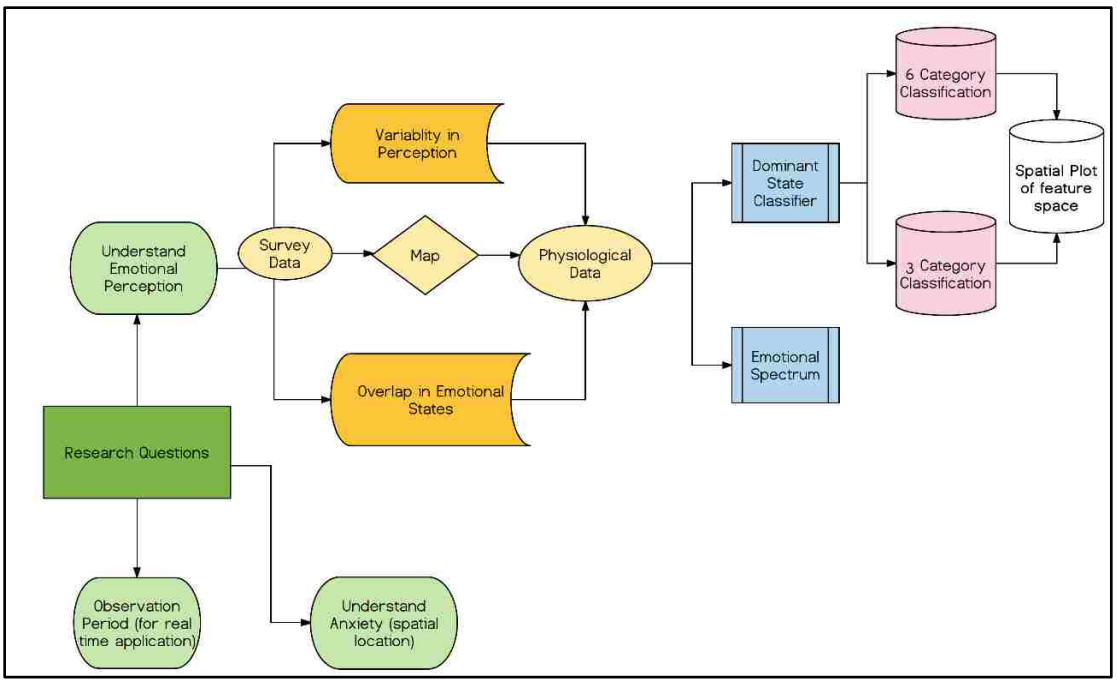


Figure 3 Pertinent Research Questions

The data for our study was collected using an application deployed on R Shiny server. Each subject watched 9 videos, 1 from each emotional state and the other 3 were randomly selected from distinct emotional states. Physiological time series data including heartrate, blood volume pressure, electrodermal activity and inter-beat interval was recorded using the Empathica E4 wearable sensor watch [23], [47]. After each video, the subjects completed a survey rating the emotional states on a scale of 1-10, with an additional score for Anxiety. All participants arrived at the study location 15 minutes prior to the scheduled start time. The participants waited in the lobby with a graduate student who explains the entire process. This allowed the emotional state to be normalized before participating in the study. The survey after each video was also intended for the same purpose with the goal being to prevent bleeding of emotional responses from one video to another.

3.2.1 Videos Used for emotion induction

The primary goal of the study is to identify the dominant emotional groups using the physiological response from an individual. Thus, we must ensure that the model is not fitted to the physiological responses from a certain video of a certain length. To achieve this, the training data (used for model generation) for each emotional state contains responses from 3 distinct videos of different lengths. This allows us to capture the features that are specific to an emotional state and not a specific video. Moreover, it adds the constraint to the feature selection process, i.e. the features used for model generation should be independent of the length of physiological time series. The video data set was manually curated by our team. The video playlist can be found at, https://www.youtube.com/playlist?list=PLjCBhI2RQVqIWKcishzr22R1ghBKRNfnL

Table 1: Videos used for inducing emotion

Movie	***Target Emotion***	***Length***
Witness	Anger	2:12
Gandhi	Anger	3:02
My Bodyguard	Anger	4:20
When Harry met Sally	Amusement	3:19
On Golden Pond	Amusement	1:26
An officer and a gentleman	Amusement	2:21
Silence of the Lambs	Fear	3:56
Halloween	Fear	4:16
Marathon Man	Fear	3:00
Pink Flamingos	Disgust	1:07
Maria's Lovers	Disgust	1:42
The Godfather	Disgust	1:53
An officer and a gentleman	Sad	2:33
The killing fields	Sad	2:30
The Champ	Sad	4:08

The Last emperor	Neutral	2:04
Hannah and her sisters	Neutral	2:16
All the Presidents' men	Neutral	2:02

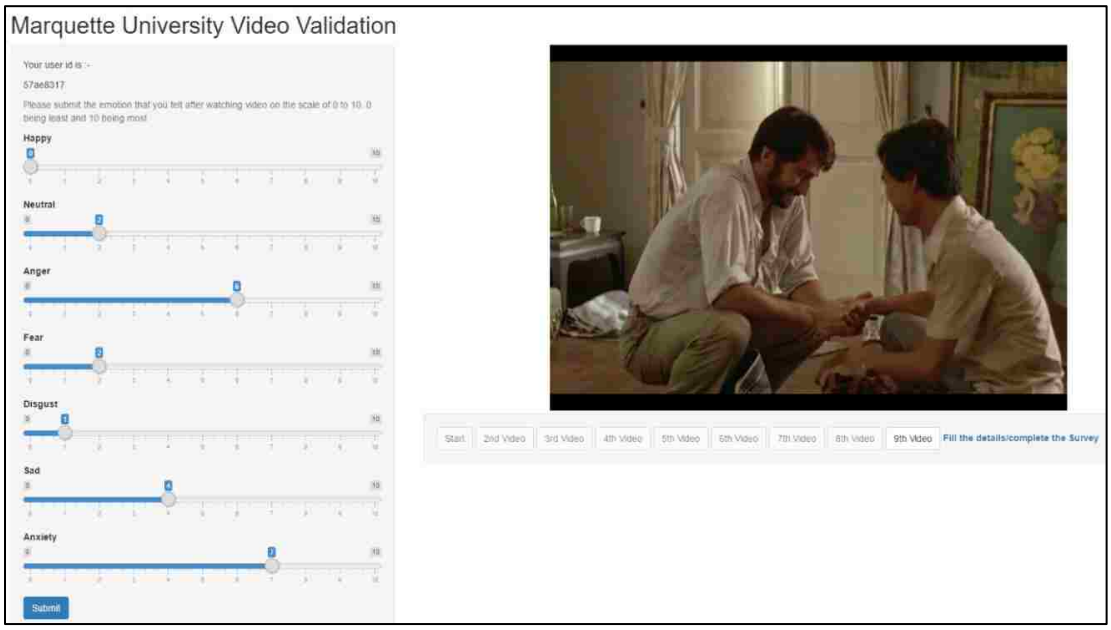


Figure 4: Data Collection application hosted on shiny.io

The data collection application can be found at https://marquetteubicomp.shinyapps.io/Validation/

3.2.2 Application architecture

A sample instance of the data collection process can be seen below in Figure 4. Each study consists of 9 instances tied together by a random ID. IRB approval was obtained by Marquette University's office of research and sponsored programs (OSRP) []. One of the defining features of this work is the size of the subject pool. The objective here was to accommodate every participant within a 30-minute window and have a less than 5-minute turnover (time between consequent participants) while maintaining data quality and integrity.

The video-survey loop is repeated 9 times. Every new participant initiates a unique randomly generated key. After each survey, a .csv file with the random key, video ID (string defined by unique video name and the order in which the video appears), time stamps (beginning and end of video) and 1x7 array from the self-report is uploaded to drop-box. Once the subject clicks the final submit button a subroutine (R script) automatically moves the data for the instance (entire study duration for a subject) from the watch to the E4 administered server and from the E4 server to drop box with the unique key generated (generated for the survey data) for the individual. The process generates a data frame in R. This allows the researchers to run data validation and integrity subroutines in real time and identify errors caused due to equipment failure or software failure (glitch in the collection application) in real time. Moreover, this data structure allows for easy analysis since the data can be sorted by individual subjects, video-id, survey results or the target class. The target class is the class associated with the video as validated in Hewig et al.[46].

Table 2: Data Structure for each instance

Subject ID	Video ID	Time Start	Time End	E4 Data-HR	E4 Data-EDA	E4 Data-IBI	E4 Data-BVP	Survey Results	Target Class
INT	STRING	NUM	NUM	TS	TS	TS	TS	1X7 ARRAY	STRING

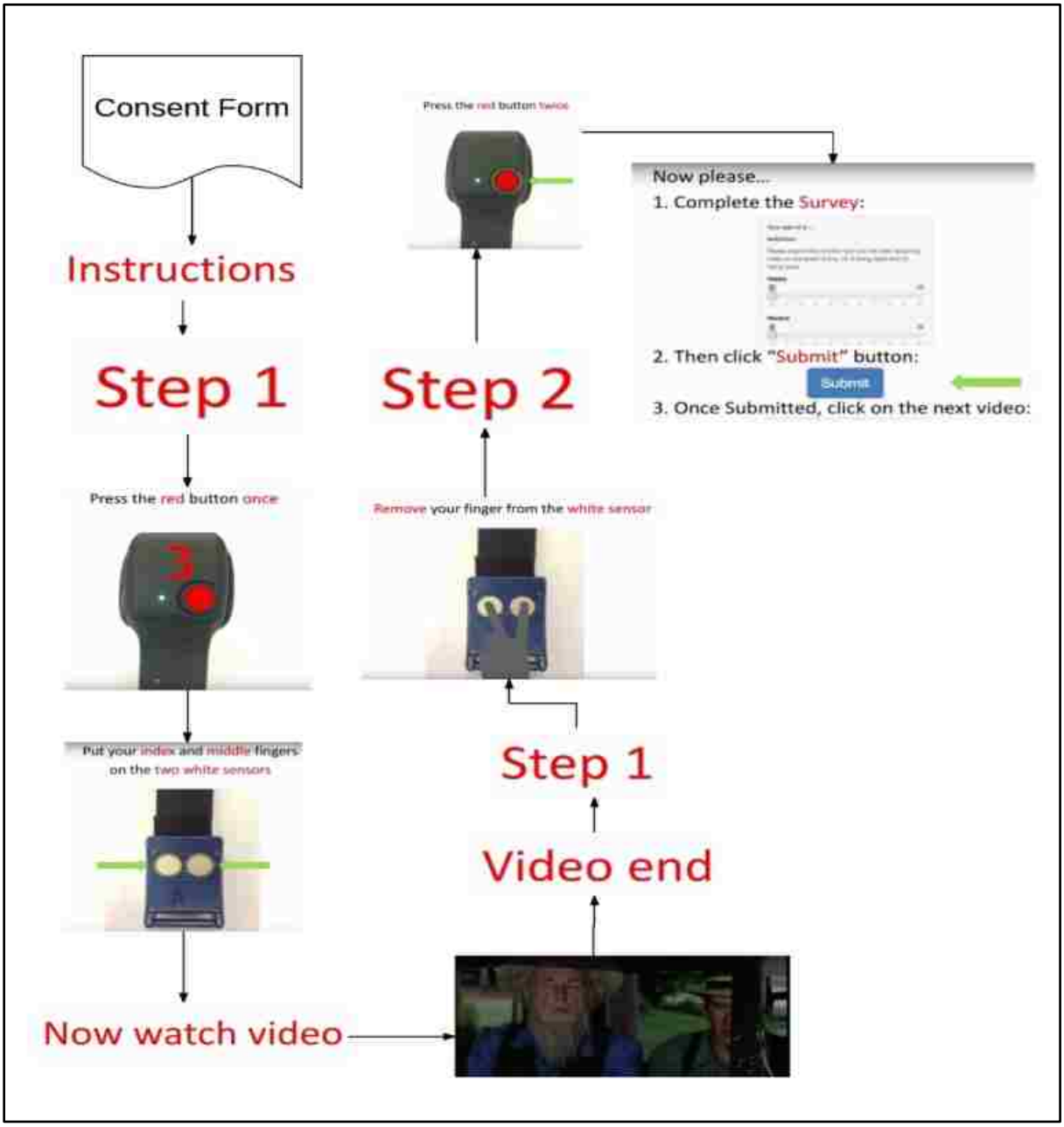


Figure 5: Data collection sample instance workflow

3.2.3 Physiological Signals

The physiological signals were recorded using Empatica's E4 wristband and Q-sensor[23], [47].

Electrodermal activity(EDA): EDA is a measurement of the changes in the skin conductance. Emotional activation, increased cognitive workload or physical exertion lead to the bodies response of sweating. The electrical conductance increases significantly (to be detected by sensors) due to the increased sweat accumulation in sub dermal pores[47]. The E4 sensor passes a small current through the electrodes in dermal contact and measures the skin conductance. Higher activation leads to larger volume of sweat accumulation in subdermal pores and thus, higher skin conductance. The EDA is measured in micro Siemens.[47] . The compound EDA signal is composed of,

Tonic EDA: This refers to the baseline skin conductance, in absence of external stimuli. Graphically, these are the smooth underlying slowly changing signals[47].

Phasic EDA: These refer to the abrupt increase in the skin conductance level. Phasic EDA is not continuous and highly correlated with external stimuli[47].

Blood Volume Pulse, inter beat interval and Heartrate: The E4 uses the Photopletysmography (PPG) to estimate the Blood volume pulse[47]. This is the same technology that is used in most modern day wearable sensor watches. Heartrate is derived from the PPG signal by computing the intervals between adjacent peaks. The inter beat interval timing is used to compute the instantaneous heart rate. The E4 watch combines a red and a green light to remove motion related artifacts from the BVP signal[47]. The IBI signal

refers to the distance between heartbeats. The algorithm to calculate the PPG is a proprietary and undisclosed[47].

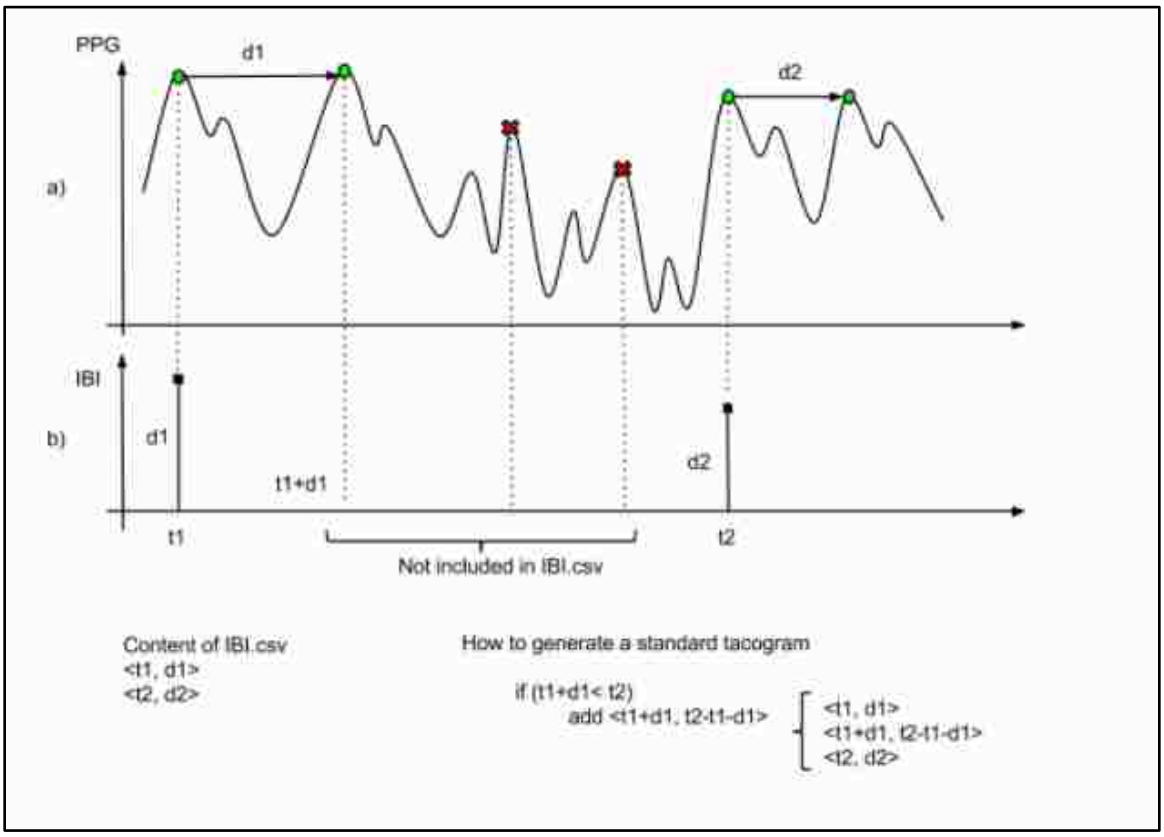


Figure 6 : IBI calculation using PPG[47]

3.2.4 Data Cleaning and Preprocessing

The biggest source of error was empty surveys, i.e. the individual watched the video but submitted a survey with zeros for all response categories. While this data can be used to see if there is a map between the physiological time series data and the true classification (based on Hewig et al.[46]) it cannot be used to model the emotional spectrum. Thus, the data

corresponding to zero survey responses was discarded. Other sources of error included equipment failures and human error. There were instances where the sensor watch recorded no data and had to be reset and instances where the individual loosened the sensor watch leading to non-continuous dermal contact and thus, erroneous readings. The application was created to ensure minimal pre-processing with subroutines that generated indicators of data quality for each instance. These are explained in detail in the experiment design section.

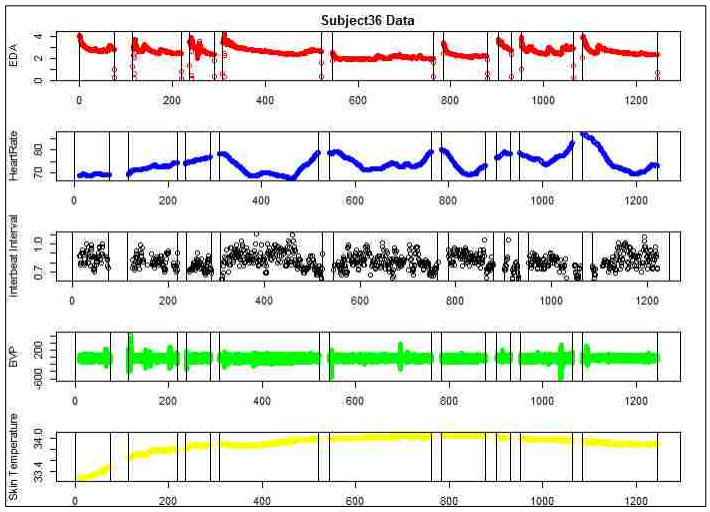


Figure 7: Splitting the Time series data by instance

3.3 Research Questions

To model emotional response as a function of physiological sensor data we need to develop a better understanding of emotional response. These questions (below) are critical for modelling the emotional spectrum and have implications in the development of a real-time application. In this section, we develop the hypothesis that will be tested in the later sections.

3.3.1 Understanding Emotional Perception

i. Emotional perception is variable and differences exist amongst individuals. While majority of the data should validate the true class (dominant emotional state) established in Hewig et al. we expect instances with deviations [46]. Two completely similar physiological states might be caused due to different emotional states in two individuals. There is a higher likelihood that variation in perception would arise in emotional states closely related e.g. anger and fear.
ii. We predict the existence of an emotional spectrum. The emotional state might have a dominant component but it is a mixture of emotions.
iii. We can leverage the physiological time series data to spatially locate the emotional states and understand their overlap. For instance, “Amusement” and “Disgust” might lie on the opposite ends of the spectrum, while “Sadness” and “Fear” might lie closer to “Disgust” and even overlap. We can reconstruct this space using the features extracted from the physiological time series that best distinguish between the target class. This feature space is unique for every individual and a function of time. It changes over time with life experiences.

3.3.2 Understanding Anxiety

The original dataset from Hewig et al. did not include anxiety as an emotional state [46]. The survey was designed to include a score for anxiety. We predict that anxiety exists as a non-continuous emotional state that overlaps multiple dominant emotional states. i.e. it is possible to be anxious waiting for good news, overlap with / proximity to “Amusement” and it is

possible to be anxious anticipating something negative, overlap with /proximity to “Fear”, “Disgust” or “Anger”

3.3.3 Existence of a map from physiological data to dominant emotional state

We predict that the existence of a unique feature space (derived from the physiological data) which can be leveraged to train a machine learning classifier that distinguishes between the target classes with very high accuracy.

3.3.4 Existence of a map from the physiological data to the survey data

We predict the existence of a map (multivariate regression or neural net) that connects the survey data to the physiological data (feature space derived from physiological data). Thus, we believe it is possible to predict survey results with high accuracy given the physiological feature space.

3.3.5 Observation Period for real-time application

The primary goal of the data analysis is to develop a model independent of the length of the physiological time series that identifies the dominant emotional state and the corresponding spectrum. However, for a real-time implementation we need to identify a time “t” for which to extract the features, classify the dominant emotional state and generate a spectrum. We suspect this would be a machine learning and simulation problem.

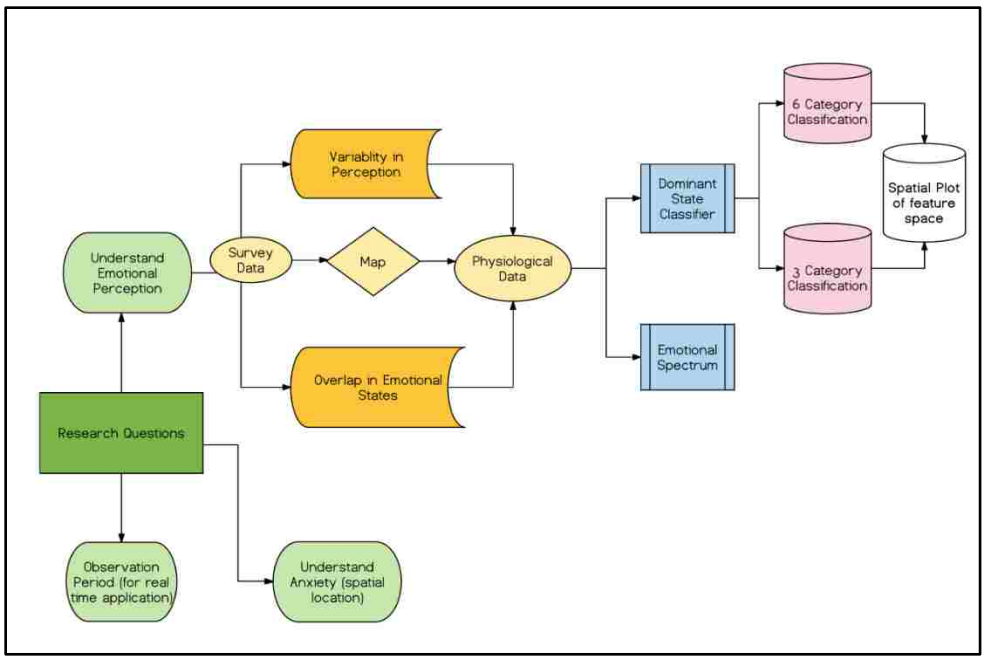


Figure 8: Pertinent Research Questions

3.4 Survey Data Analysis

As we test the hypothesis and try to get answers to the research challenges posed it is important to identify the end goal. We want to use physiological sensor data to identify the dominant emotional state and predict the emotional spectrum. We expect higher error rates while predicting the emotional spectrum due to variability in emotional perception. However, we expect both models to serve as a basis for the iterative learning that allows for the model to evolve and tailor itself to the emotional perception of the individual. We will use survey responses to model the emotional spectrum as a function of the physiological time series features.

3.4.1 Survey Responses

We would expect the distribution of the survey results for a target emotional group to be unimodal. i.e. for a target class (Amusement, Anger, Disgust, Sad, Fear, and Neutral) all responses would lie within that target class. However, this is found to be not true. Table 3 contains correlations amongst columns for each of the target class. The data from each subset was normalized (mean of column set to 0 and variance set to 1). The reason for this step is to have a relative distribution of the column (survey responses for the magnitude of the state) across the entire population. The p-values associated with the correlations can be found in Table *4*.

Table 3 Correlations between dominates states for a target class

Target Class	**Survey Reponses-Correlation**						
	Amusement	Neutral	Anger	Fear	Disgust	Sad	Anxiety
Neutral	0.088561	1	-0.10399	0.15654	-0.02272	-028448	0.009395
Anger	0.094521	-0.01167	1	0.330934	0.414545	0.46673	0.293252
Amusement	1	-0.20043	0.174053	-0.01881	-0.12139	0.190844	-0.01241
Fear	-0.07152	-0.19728	0.187189	1	0.100622	0.198019	0.623399
Disgust	-0.29017	-0.30572	0.466622	0.46505	1	0.385963	0.482161
Sad	0.062531	-0.31094	0.247254	0.39548	0.229376	1	0.336707

Table 4 P-value for the correlations in Table 3

Target Class	**Survey Responses**						
	Amusement	Neutral	Anger	Fear	Disgust	Sad	Anxiety
Neutral	0.395989	NA	0.318571	0.131884	0.82794	0.005455	0.928389
Anger	0.383848	0.914553	NA	0.001743	6.56E-05	5.20E-06	0.005841
Amusement	NA	0.059661	0.102832	0.86111	0.257122	0.073217	0.908082
Fear	0.502937	0.062351	0.077286	NA	0.345353	0.061361	5.30E-11
Disgust	0.006406	0.003982	5.23E-06	5.67E-06	NA	0.000222	2.26E-06
Sad	0.55821	0.002855	0.0188	0.000114	0.029652	NA	0.001175

3.4.2 Variability in emotional perception

Emotional perception is unique to everyone. Despite being in the same physiological state the perception can be different amongst two individuals. The boxplot below shows the distribution of level of Neutral (survey response) across all dominant emotions. While most of the survey responses lie within the neutral category, there are statistically significant responses under disgust and amusement indicating, the subjects had a neutral response to videos intended to induce amusement and disgust. A possible reason for this could be cultural changes. The videos used in the study are from popular movies released before the year 2000 while the study population consists of individuals between the ages of 18 and 21. This it is likely, that certain videos failed to induce the target emotion in a portion of the population. This is different from overlap of emotional states (discussed in 3.4.3) since there

is almost no correlation between “Neutral” target class and “Amusement”, “Disgust” survey responses. Moreover, Amusement and Disgust lie on opposite sides of Neutral with respect to emotional perception. Thus, we can conclude that this is an example of the variability in emotional perception.

Table 5: Correlations amongst survey responses for "Neutral"

Target Class	**Survey Reponses-Correlation**						
	Amusement	Neutral	Anger	Fear	Disgust	Sad	Anxiety
Neutral	0.088561	1	-0.10399	0.15654	-0.02272	-028448	0.009395
Target Class	**Survey Responses, P-value associated with the correlations**						
	Amusement	Neutral	Anger	Fear	Disgust	Sad	Anxiety
Neutral	0.395989	NA	0.318571	0.131884	0.82794	0.005455	0.928389

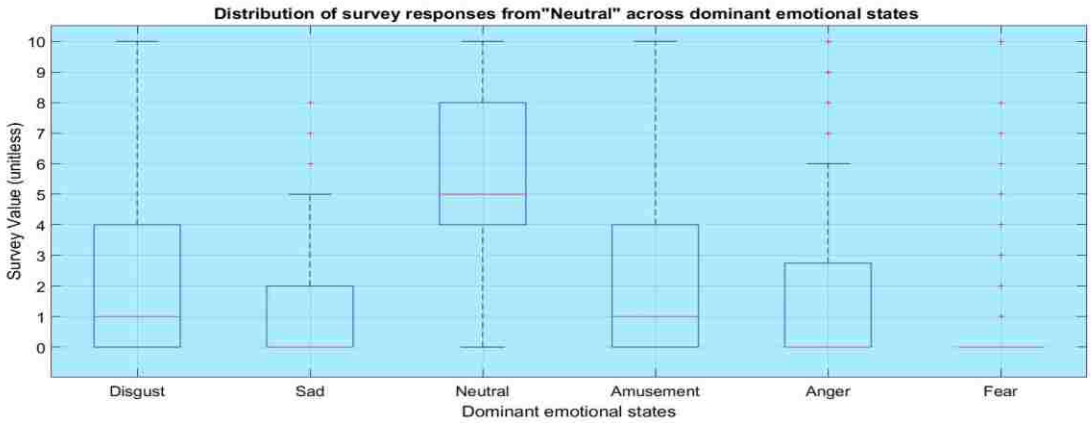


Figure 9: Distribution of survey responses from “Neutral" across dominant emotional states

3.4.3 Overlap in emotional states

We must also account for the overlap of emotional groups. While there is no overlap in extreme states such as amusement and disgust, there is significant overlap between closely associated emotions. The boxplot below shows the distribution of level of Disgust (survey response) across all dominant emotions. Unlike Table *5*, where there is a statistically significant difference in the mean of responses between Neutral and other emotional states the, box plot in Figure *9* indicates an overlap between the states of Disgust (target class), Anger and Fear. Moreover, Fear and Anger survey responses have a relatively high, statistically significant positive correlation for the target class "Disgust".

Table 6: Correlations amongst survey responses for "Disgust"

Target Class	**Survey Reponses-Correlation**						
	Amusement	Neutral	Anger	Fear	Disgust	Sad	Anxiety
Disgust	-0.29017	-0.30572	0.466622	0.46505	1	0.385963	0.482161
Target Class	**Survey Responses-, P-value associated with the correlations**						
	Amusement	Neutral	Anger	Fear	Disgust	Sad	Anxiety
Disgust	0.006406	0.003982	5.23E-06	5.67E-06	NA	0.000222	2.26E-06

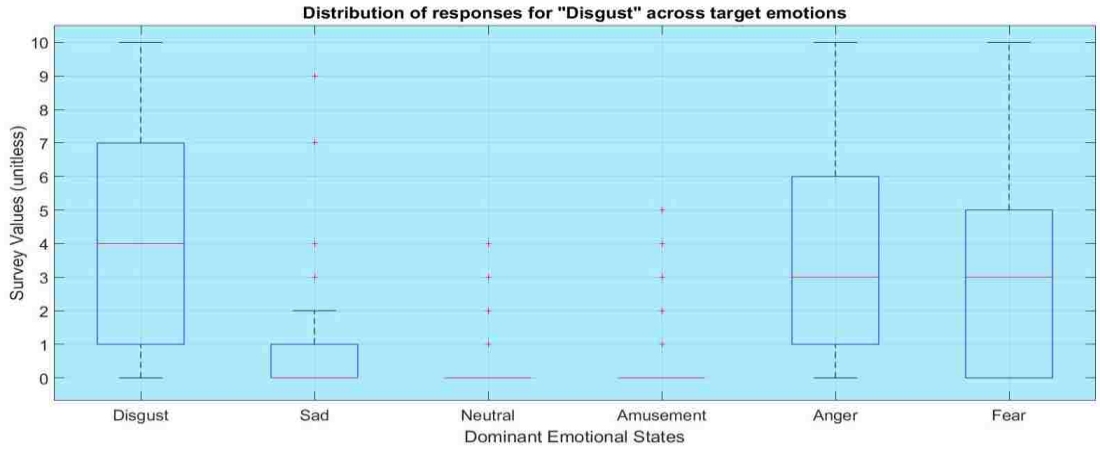


Figure 10: Distribution of responses for "Disgust" across target emotions

3.4.4 Anxiety

The boxplot (Figure *11*) below shows the distribution of level of Anxiety (survey response) across all dominant emotions. The survey responses indicate that anxiety is a discontinuous state that intersects with the emotional states of disgust, sad, anger and Fear. The means and the standard deviations are statistically very close for the states of disgust, sad and anger while the mean is higher for the state of fear, indicating that higher levels of anxiety occur in conjunction with fear. Moreover, there is statistically significant correlation between the target class's "Fear", "Disgust", "Anger", "Sad" and survey responses for "Anxiety".

Table 7: Correlation between Anxiety responses and responses from target class's

	Dominant Emotional States (survey responses)					
Anxiety	Amusement	Anger	Fear	Disgust	Sad	Neutral
Correlation	-0.012	0.293	0.6233	0.4821	0.3367	0.009
P-value	0.928	0.0058	0.0000	0.0000	.001	0.092

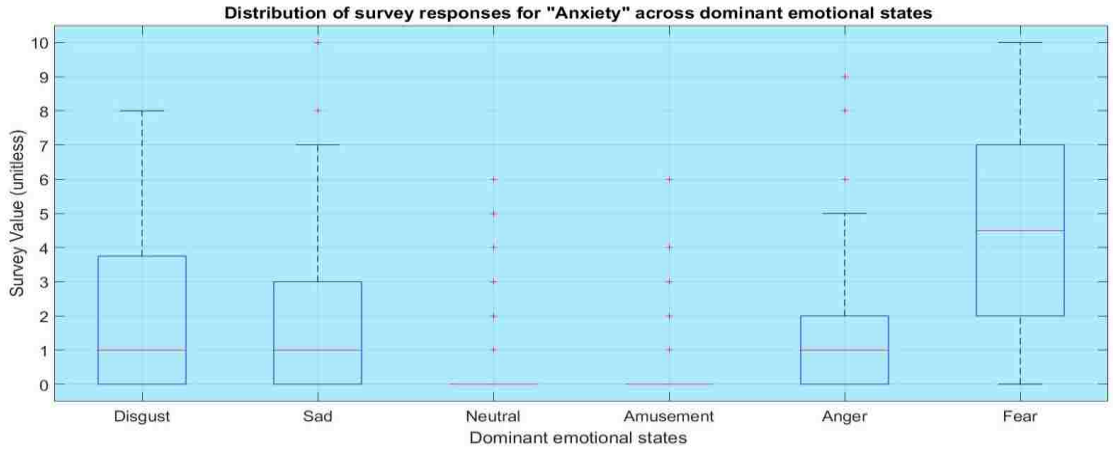


Figure 11: Distribution of "Anxiety “across dominant emotional states

3.4.5 Predicting Dominant State using Survey Data

Validating survey results provides us an insight into individual emotion perception, overlaps in emotional states and establishes the credibility of the data set as a reliable source of audio-visual stimuli for induction of the target emotional state. For this model, we used 537 survey observations of 6 dominant emotional states. We used a Linear SVM classifier. While the

model achieves an overall accuracy of 73.7 %(25%hold-out validation), it provides important insights into emotional perception. The model results indicate,

i. The variability in emotion perception and overlap between emotional states is more significant between states that are closer (we will define a feature space later in this paper). E.g. 21 % of videos with "Disgust" as the true response were miss-classified as "Fear. This reinforces the concept of overlap of emotional states discusses in 3.4.3.

ii. The neutral category has the highest false discovery rate. This reinforces the concept of variability in emotional perception discussed in 3.4.2.

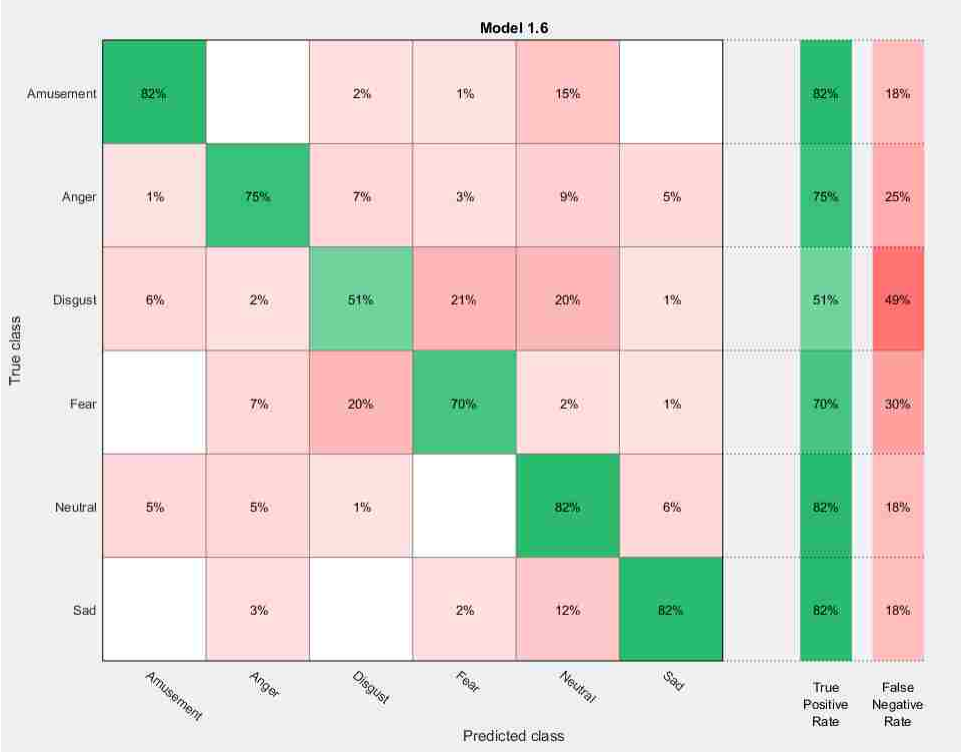


Figure 12: Confusion Matrix -predicting dominant emotional state given survey data

Figure 13: Confusion Matrix 2- predicting dominant emotional state given survey data

3.5 Predicting Dominant Emotional State -Six Response Classes

The next step per the research objectives is to find a map between the physiological data and the dominant emotional state. To accomplish this, we must assume there exists a feature space that explains the variability within physiological data for our target class's. Once this feature space is found we train multiple machine learning classifiers to identify which classifier leads to the highest accuracy for the chosen feature space.

3.5.1 Feature Selection

The feature selection process was ad-hoc. We examined each possible feature individually to determine if its use as a classifier would be warranted. One again due to variable length time series, the features was scale independent. Our analysis revealed that the RMS levels (for EDA, HR, and BVP) were the best predictors for classification into the 6 dominant emotional states. We tested feature spaces composed of multiple predictors from time domain and the outlined in [34]. These include, mean, median, variance, mean of absolute first differences, mean of absolute second differences, mean of absolute value of first differences and mean of absolute value of second differences. The frequency domain signals included magnitude and phase information from signal FFT, signal periodicities, signal power etc. The RMS level of normalized signals performed best with our dataset. We also tried a reconstructed phase space approach well suited for non-linear time series[48], [49]. While the approach shows promise the classification accuracy was lower than the one through RMS level and the computational needs are significantly higher making real-time implementation challenging.

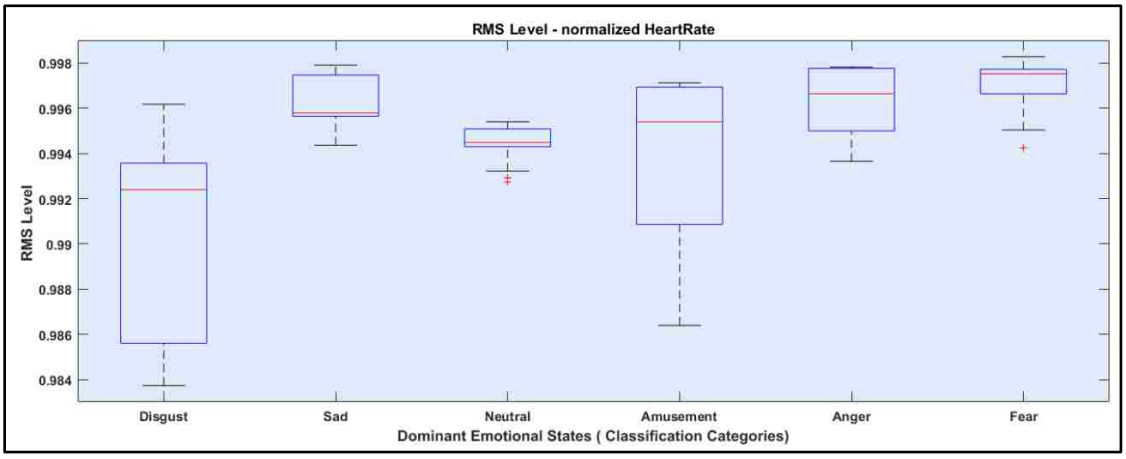


Figure 14: Distribution of HR feature across 6 dominant states

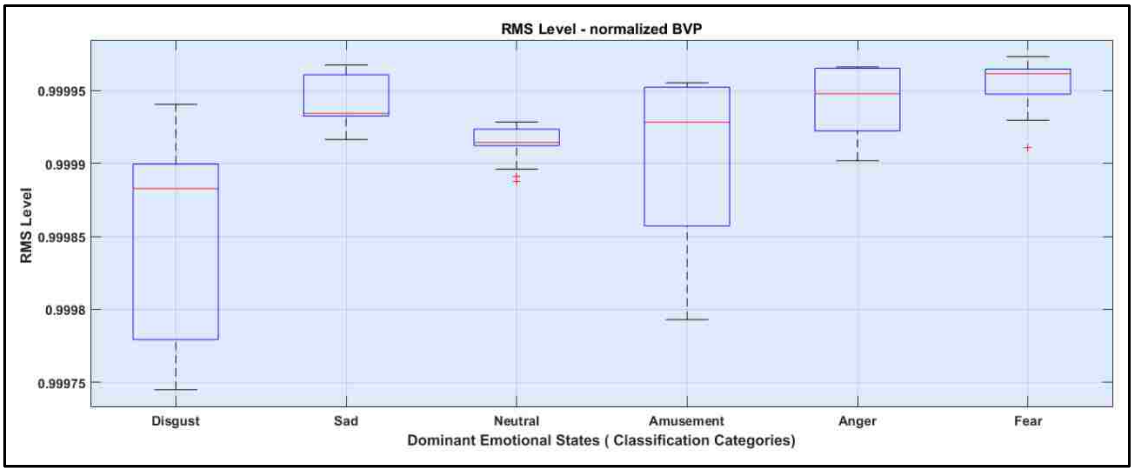


Figure 15:Distribution of BVP feature across 6 dominant states

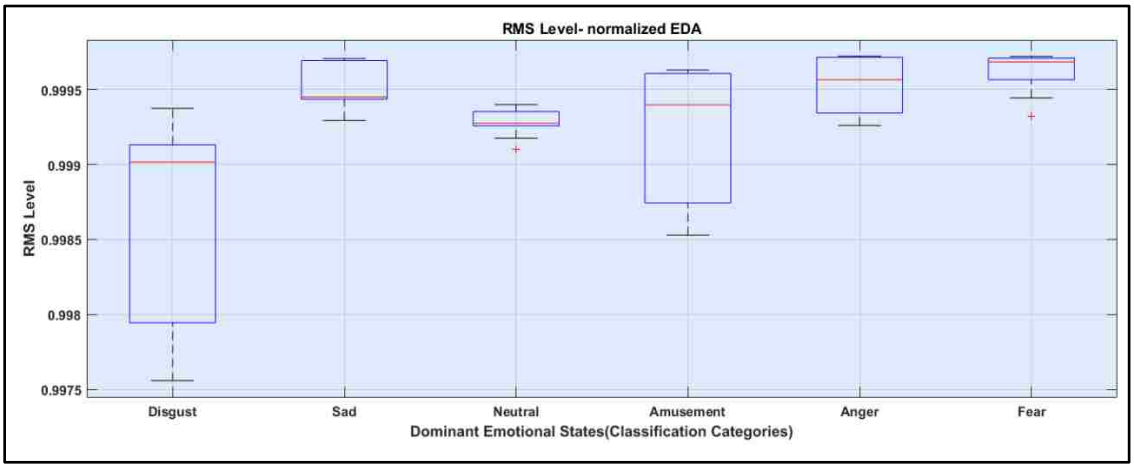


Figure 16: Distribution of EDA feature across 6 dominant states

3.5.2 Prediction accuracy for a 6-category classification

The feature boxplots in Figure *14*,Figure *15* and Figure *16*validate the use of root mean square level of the physiological data (hear rate, electrodermal activity and blood volume pressure) as predictors for the classification model. The model was trained using 410 instances and tested

on 140 instances. The cubic KNN model achieves the highest accuracy (using only physiological data) using all three features (87%). Using heartrate alone the cubic KNN achieves an accuracy of 77%. A cubic SVM that combined the physiological data with the survey responses achieves an accuracy of 93%. This is significantly higher than the accuracy achieved using survey responses alone (79%). While this, does not have any implications for the real-time application (uses only physiological data), it reiterates the information gain due to the physiological data. We calculate RMS for normalized signals.

$$X_{rms} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} |X_n|}$$

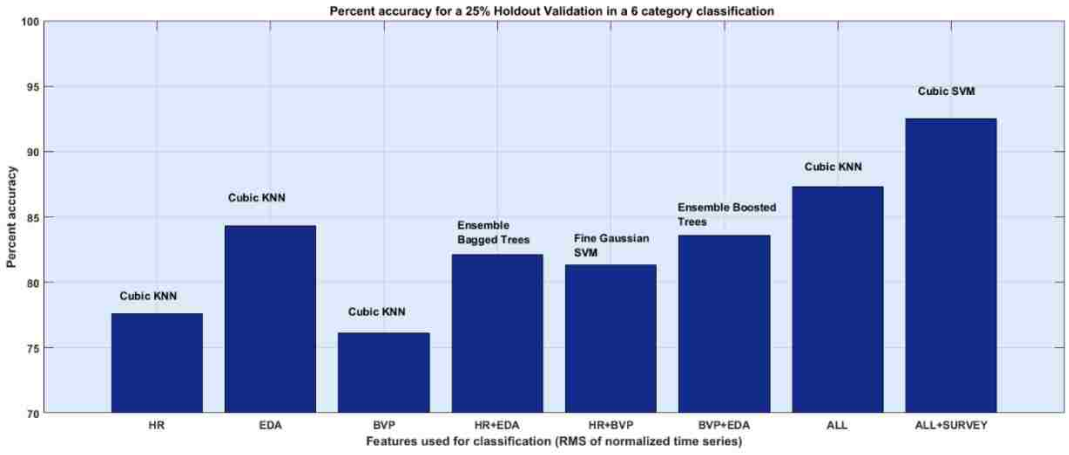


Figure 17: Model Accuracy for a 6-category classification

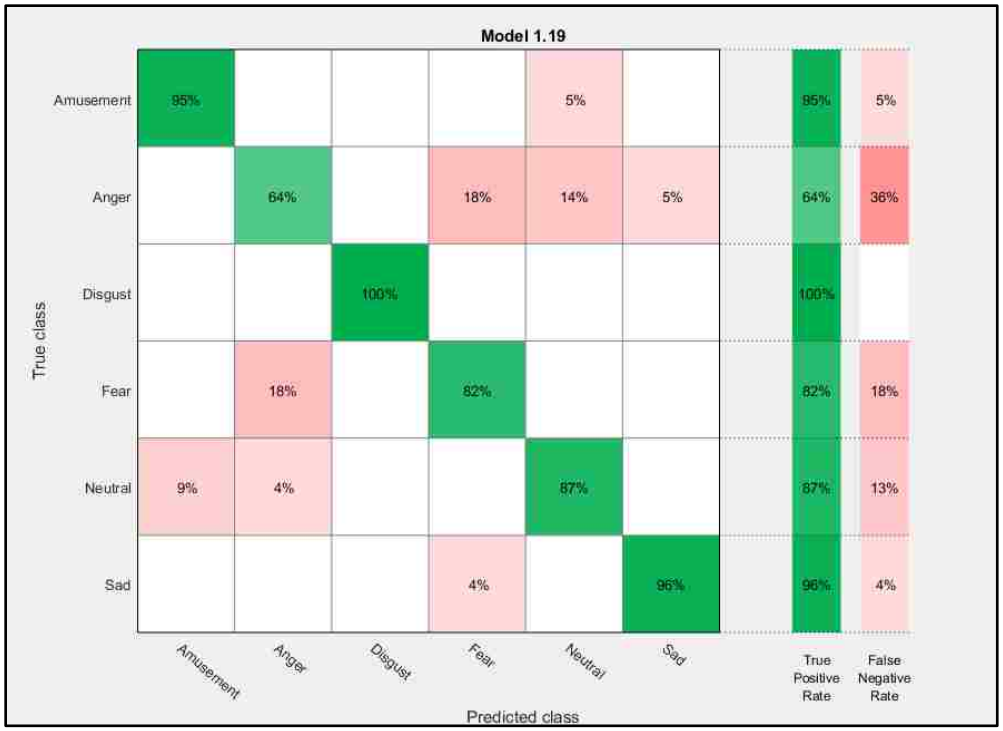


Figure 18: Confusion matrix for 6-category classification | 87.4% accuracy

An accuracy of 87.4 % was achieved using the feature space constituted by EDA, BVP and HR time series features using a Cubic KNN classifier. This result further validates the concepts of variability in emotional perception and overlapping emotional states. The accuracy of physiological data – target class is higher than survey data- target class. This indicates similar physiological states in two individuals might have different emotional states. In most scenarios, we expect variability in perception, for instance “Neutral” being classified as “Amusement” in some cases and “Anger” in some cases. While, in other cases we can expect an overlap of dominant states. Such as Fear being classified as “Anger” and vice versa.

3.5.3 Feature Space for a 6-category classification

As hypothesized in 3.4.1 we plot the feature space for the six dominant emotional states. The observations (features representing the dominant emotional states) are represented by points in the plot using principle component analysis. The feature space validates the hypotheses in 3.4. The cluster centers (center of the ellipse) for Disgust and Amusement are far apart. There is also significant overlap between Fear, Sad and Anger. Figure *19* is a PCA visualization of the feature spaced used to train the classifier. It makes sense intuitively since the cluster centers of all negative emotional states except Disgust are a lot closer and show significant overlap while the distance of the "Neutral" cluster is closer to Amusement than Disgust.

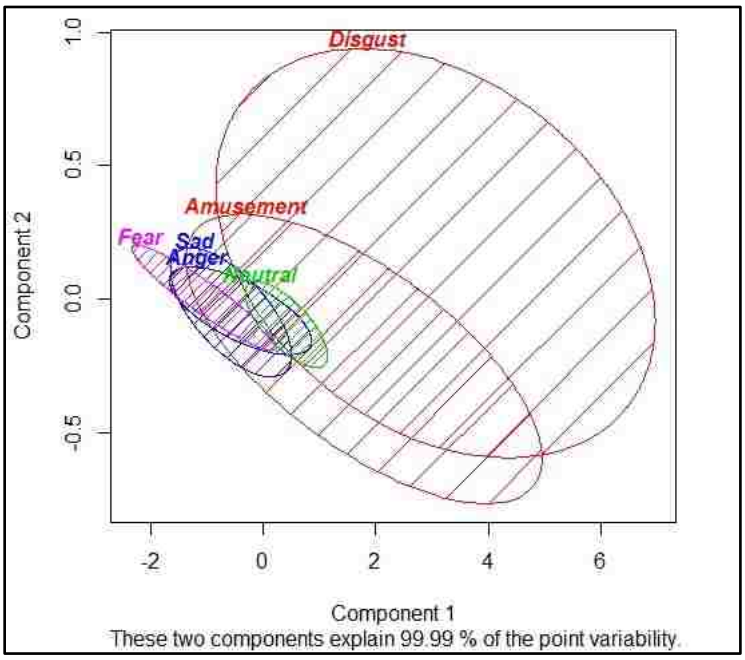


Figure 19: Spatial Locations of Emotional States for a 6-category classification

3.5.4 KNN Classifier

A KNN Classifier is used with the following hyper parameters,

- The number of nearest neighbors , 5
- Distance measure , Minkowski
- Weight measure , equal
- Distance weight measure , inverse square

These hyper parameters are validated in the Evaluation section. The cubic KNN classifier is one of the simplest classifiers. It is very effective for our problem given the low dimension feature space.

3.6 Predicting Dominant Emotional State -Three Response Classes

For a three-response classification, the data from Sad, Angry, Disgust and Fear was binned together in a category named “Negative”. The data from Amusement was named “Positive”. A data set with 90 instances of each category was randomly sampled. Statistical and time series features were extracted from each instance.

3.6.1 Feature Selection

The feature selection process was ad-hoc. We examined each possible feature individually to determine if its use as a classifier would be warranted. One again due to variable length time series, the features was scale independent. Our analysis revealed that the RMS levels (for EDA, HR, and BVP) were the best predictors for classification into the 6 dominant emotional states. We tested feature spaces composed of multiple predictors from time domain and the outlined in [34]. These include, mean, median, variance, mean of absolute first differences, mean of absolute second differences, mean of absolute value of first differences and mean of absolute value of second differences. The frequency domain signals included magnitude and phase information from signal FFT, signal periodicities, signal power etc. The RMS level of normalized signals performed best with our dataset. We also tried a reconstructed phase space approach well suited for non-linear time series[48], [49]. While the approach shows promise the classification accuracy was lower than the one through RMS level and the computational needs are significantly higher making real-time implementation challenging.

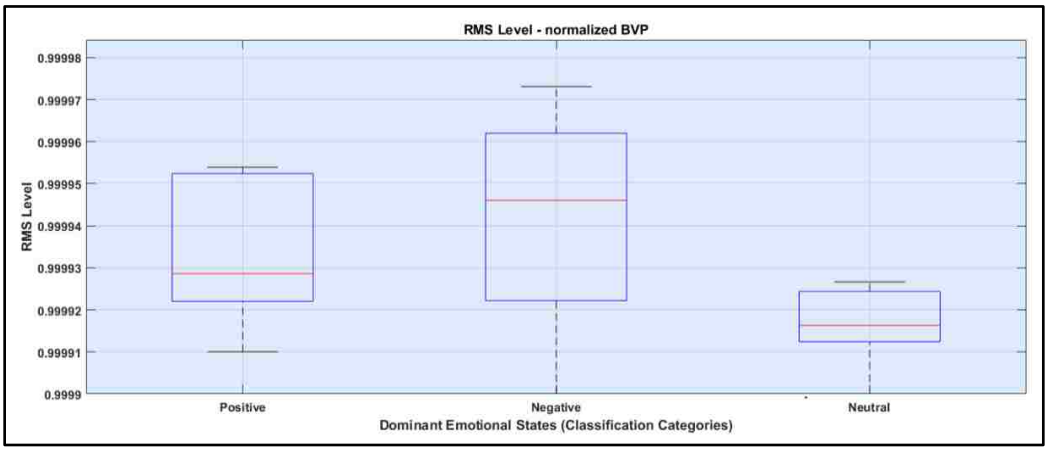


Figure 20: Distribution of BVP features across 3 emotional classes

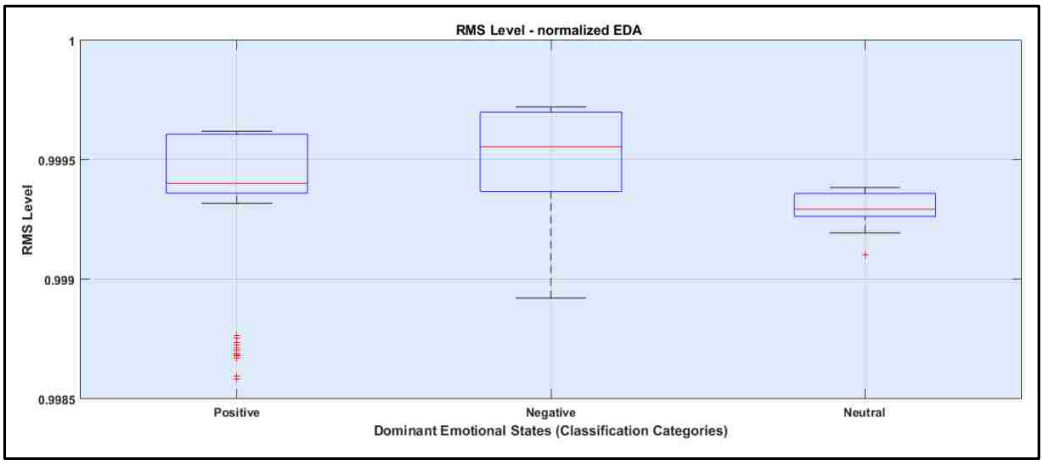


Figure 21: Distribution of EDA features across 3 emotional classes

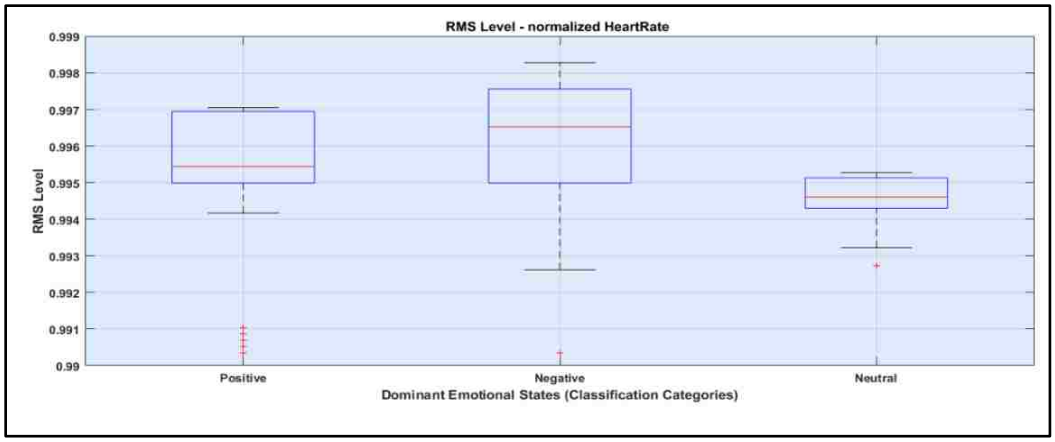


Figure 22:Distribution of HR features across 3 emotional classes

3.6.2 Prediction accuracy for a 3-category classification

The feature boxplots in Figure *20*,Figure *21* and Figure *22*validate the use of root mean square level of the physiological data (hear rate, electrodermal activity and blood volume pressure) as predictors for the classification model. The model was trained using 200 instances and tested on 70 instances. For a three-category classification (Positive, Negative and Neutral) the maximum accuracy (92%) was achieved using a weighted KNN model and electrodermal activity as a predictor. There was no information gain when heartrate and blood volume pressure were added as predictors. The weighted KNN yielded an accuracy of 87% with just heart rate as a predictor.

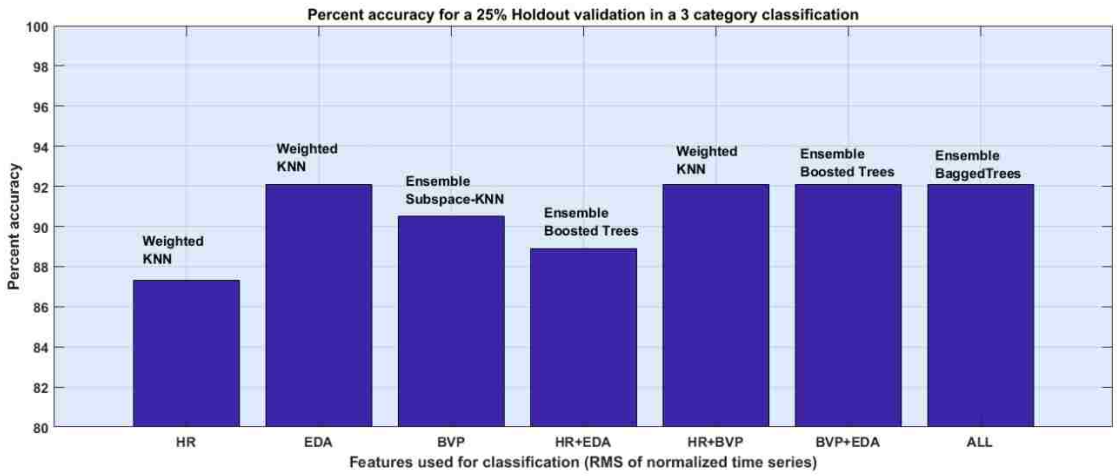


Figure 23:Model accuracy for a 3-category classification

This is a significant finding. The accuracy for a 3-category classification did not increase when EDA feature was added as a predictor. The cost of measuring BVP is significantly less compared to the upfront cost and the maintenance cost associated with EDA sensors. There is almost a 25 x cost differential and the EDA wearable sensors require electrodes be replaced after a specified period of time.

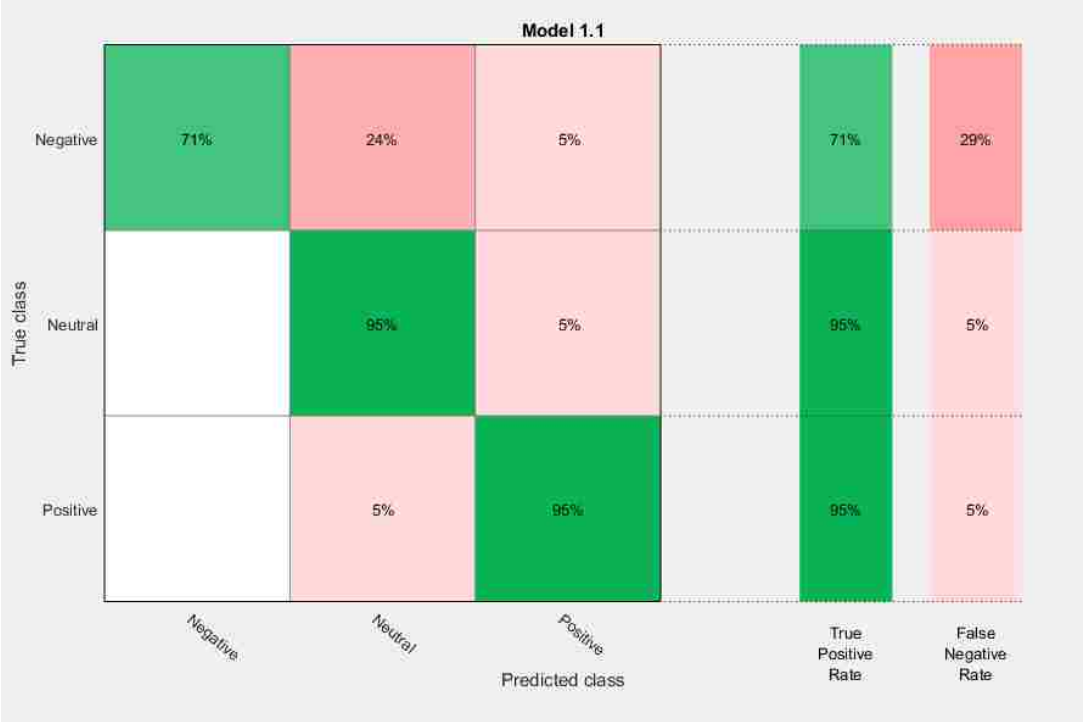


Figure 24: Confusion matrix for a 3-category classification | 92% accuracy

As expected we achieve higher classification accuracy in a 3-category classification. The concept of variability in emotion perception discussed in 3.4.2 is evident here, where in certain instances with significant “Negative” responses being classified as” Neutral”

3.6.3 Feature space for a 3-category classification

As hypothesized we plot the feature space for the three dominant emotional states. The observations (features representing the dominant emotional states) are represented by points in the plot using principle component analysis. The feature space validates the hypotheses in 3.4. The cluster centers are significantly far apart for the Negative and Positive states. The

Neutral cluster center is closer to the Positive cluster. Once again, this plot reiterates the overlap between the emotional states. Figure *25* is a PCA visualization of the feature spaced used to train the classifier. It makes sense intuitively since the cluster centers of all negative emotional states except Disgust are a lot closer and show significant overlap while the distance of the "Neutral" cluster is closer to Amusement than Disgust.

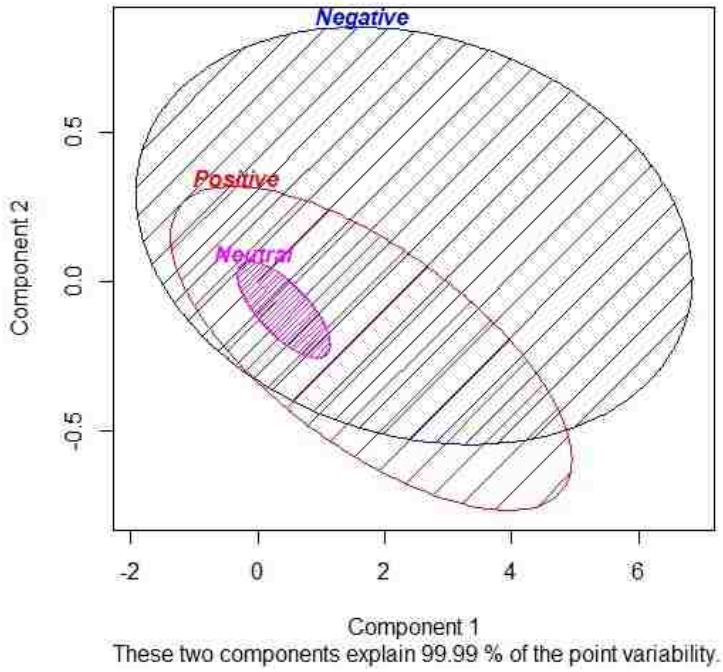


Figure 25: Spatial locations for emotions for a 3-category classification

3.6.4 Cubic KNN Classifier

A KNN Classifier is used with the following hyper parameters,

- The number of nearest neighbors , 2
- Distance measure , Euclidean
- Weight measure , equal
- Distance weight measure , equal

These hyper parameters are validated in the Evaluation section. The cubic KNN classifier is one of the simplest classifiers. It is very effective for our problem given the low dimension feature space

3.7 Predicting the Emotional Spectrum

Emotional spectrum is a 1x (n+1) that defines the activation associated with each of the n dominant emotions and Anxiety. This is a novel contribution of this work. To create an emotional spectrum, we must assume that a mapping from the physiological time series data to the survey data exists (a map between the feature space occupied by the physiological data and the feature space occupied by the survey values). In the following sections, we will present results from a 1x7 emotional spectrum and 1x4 emotional spectrum.

3.7.1 Six-Category Classification | 1x7 Spectrum

A 100 Neural network with 100 hidden neurons was used to train on the Input Data (1x3 physiological feature space and the dominant emotion using the Levenberg-Marquardt algorithm [50]-[52]. The target was set to a 1x7 array that represents activation levels for “Amusement”, “Anger”, “Neutral”, “Sad”, “Disgust”, “Fear” and “Anxiety. A 70-15-15 split between the training (375 instances), testing (81 instances) and validation (81 instances) set was used. The results from the regression are presented in Figure *27* and Figure *28*.

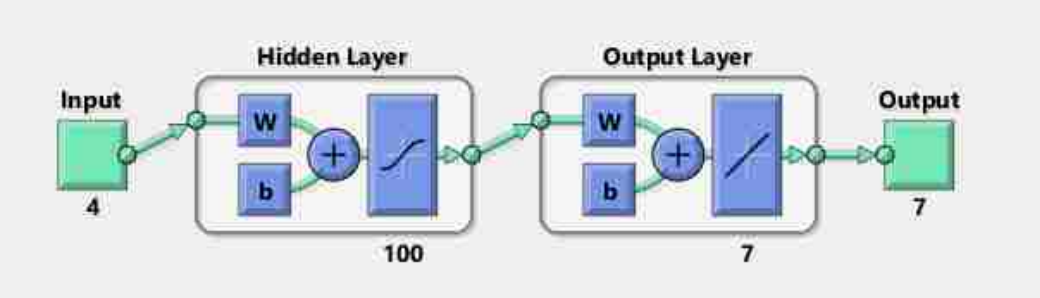


Figure 26: Network Architecture | Prediction 1x7 emotional spectrum

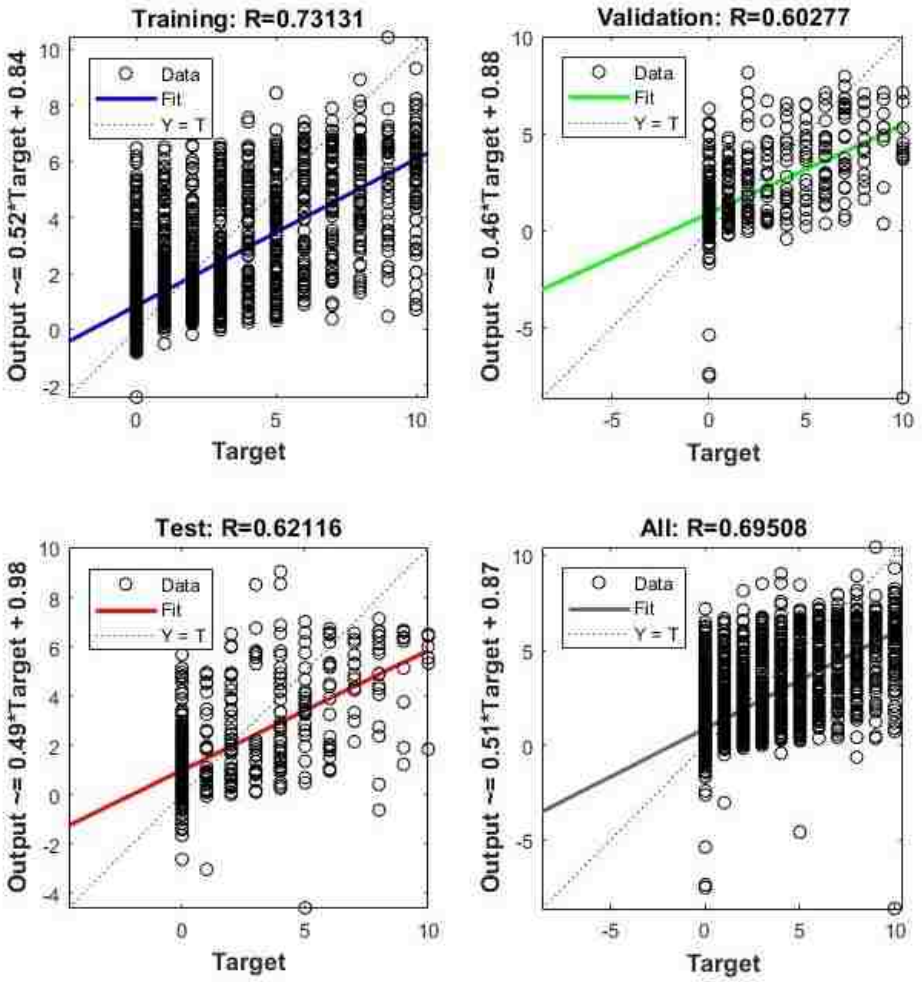


Figure 27: Regression Output | 1x7 spectrum

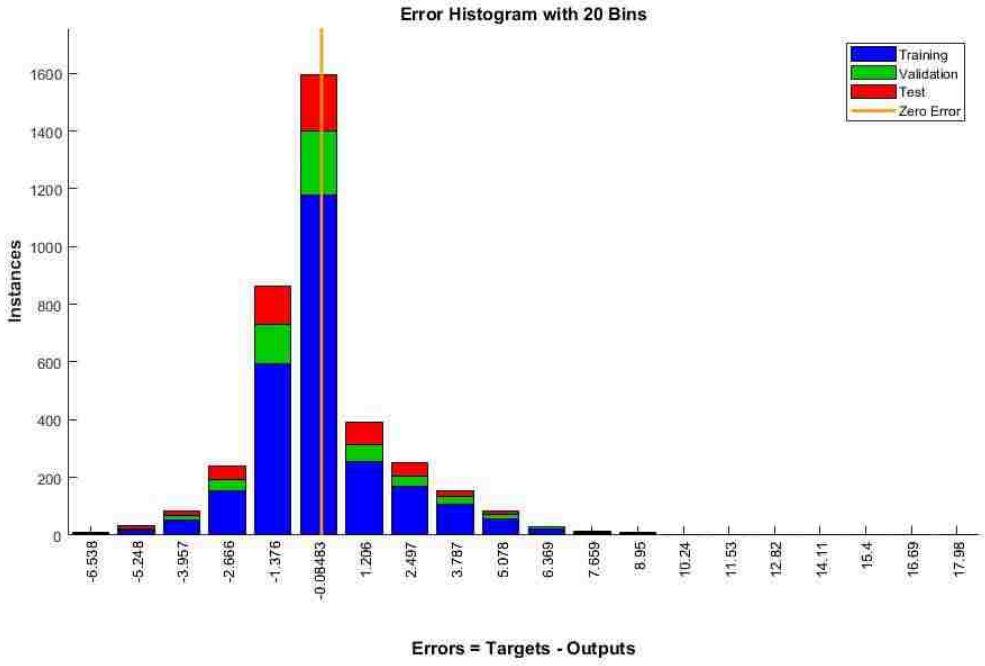


Figure 28| Error Histogram| 1x7 spectrum

3.7.2 Three-Category Classification | 1x4 Spectrum

A 100 Neural network with 100 hidden neurons was used to train on the Input Data (1x3 physiological feature space and the dominant emotion using the Levenberg-Marquardt algorithm[50]-[52]. The target was set to a 1x7 array that represents activation levels for "Positive", "Negative", "Neutral", and "Anxiety. The score for the "Negative input was computed as the mean of the nonzero scores for "Anger", "Sad", "Disgust" and "Fear". A 70-15-15 split between the training (375 instances), testing (81 instances) and validation (81 instances) set was used. The regression results are presented in Figure *29* and Figure *30*.

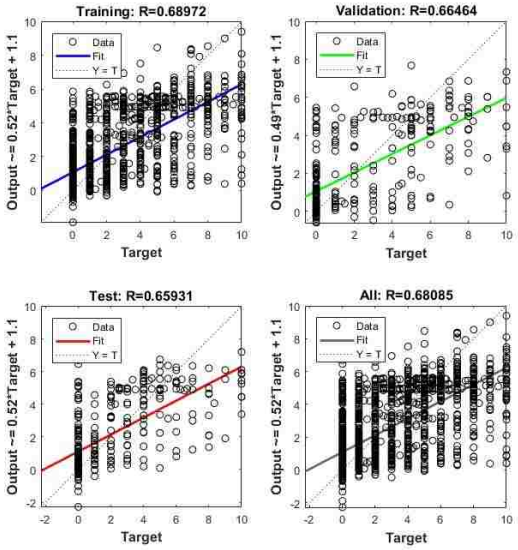


Figure 29| Regression output | 1x4 spectrum

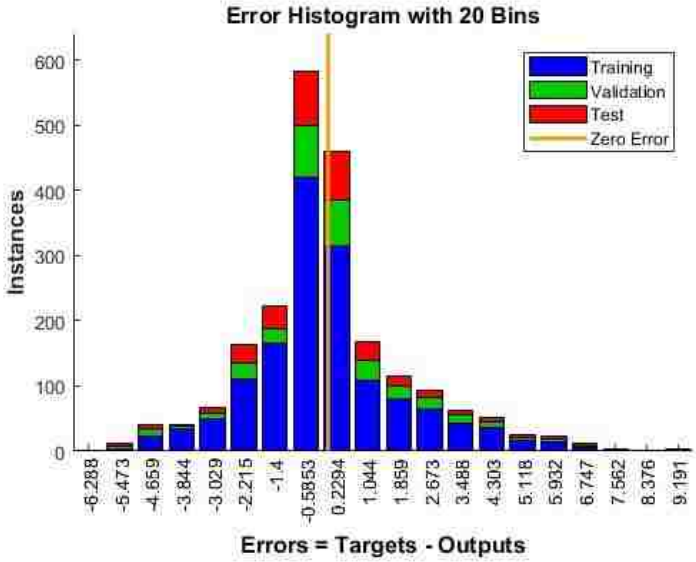


Figure 30| Error Histogram |1x4 spectrum

Regression R Values measure the correlation between outputs and targets. An R value of means a close relationship, 0 a random relationship.

3.8 Real Time Application

We have implemented the models generated for detecting the dominant emotional state and the corresponding emotional spectrum in a real-time application. We are currently running a beta test with 10 participants. The participants are graduate students in computational math. The application detects the dominant emotional state and the emotional spectrum using 60 second streams of data. We recognize that the models were trained based on the data obtained from a small portion of the demographic; moreover, there is variation in emotion perception within that population. To account this variation and the experiment population, we implemented an iterative learning framework that allows for the model to be tailored to an individual's unique emotion perception. The user data is streamed to a Matlab application that hosts the classification model developed in Chapter 3. The prediction made by model is transferred to an iterative learning module where the user input is used to modify the feature space (from the original dataset) and a new model is learned based on the modified feature space. The next prediction is made using the retrained model. The process of validation and retraining creates a model unique to the individual's emotion perception.

3.5.2 System Architecture

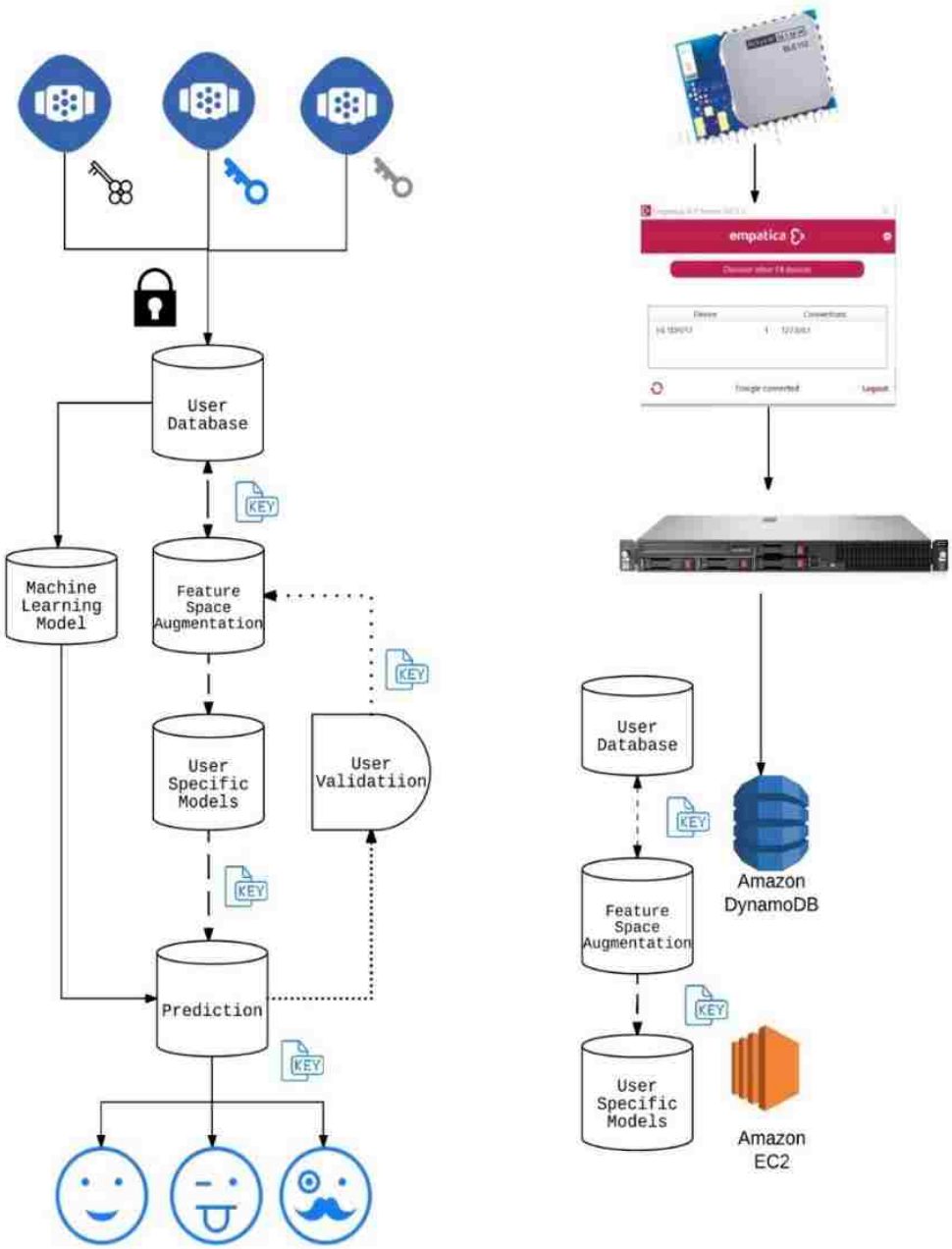


Figure 31: System Architecture

3.5.3 Iterative Learning framework

The iterative learning framework (longitudinal feature space augmentation) allows the application to improve the accuracy of predictions by tailoring the model to an individual. The dominant state and the emotional spectrum are predicted using the features extracted from physiological sensor data. User validation allows the application to modify the feature space for the individual and retrain the individual specific model.

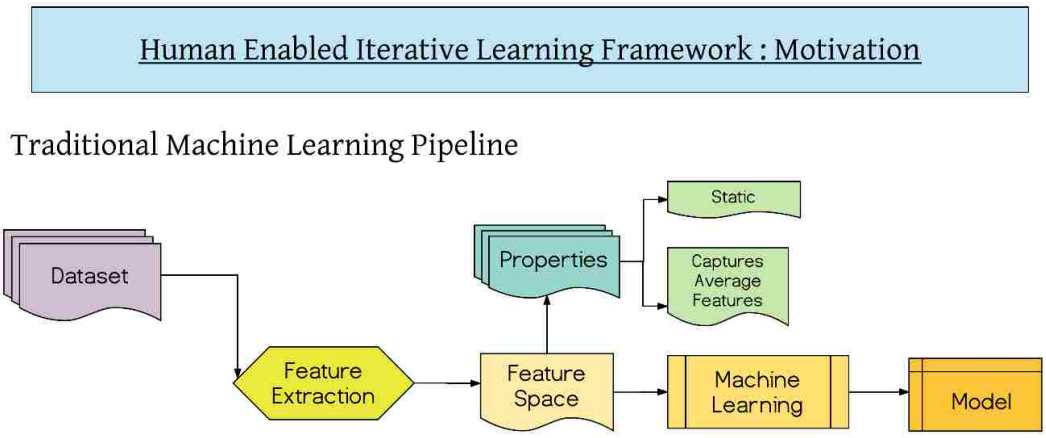


Figure 32: Motivation for Iterative Learning

The model will not do well in a highly variable feature space - predicted class problem (The distribution of feature| Response will have high variance). An example of such a situation would be prediction of emotional response. There is tremendous variability in physiologic responses (same heart rates in 2 people could be caused by opposing stimuli) and emotion perception amongst humans. The amount of data needed to account for such variability and make the model reliable would be impractical. Moreover, the feature space in the diagram above is derived from a dataset that assumes that a stimulus would induce the target

emotional response. While this may show promising results within a specific population. The accuracy will decrease when the properties of the data set change.

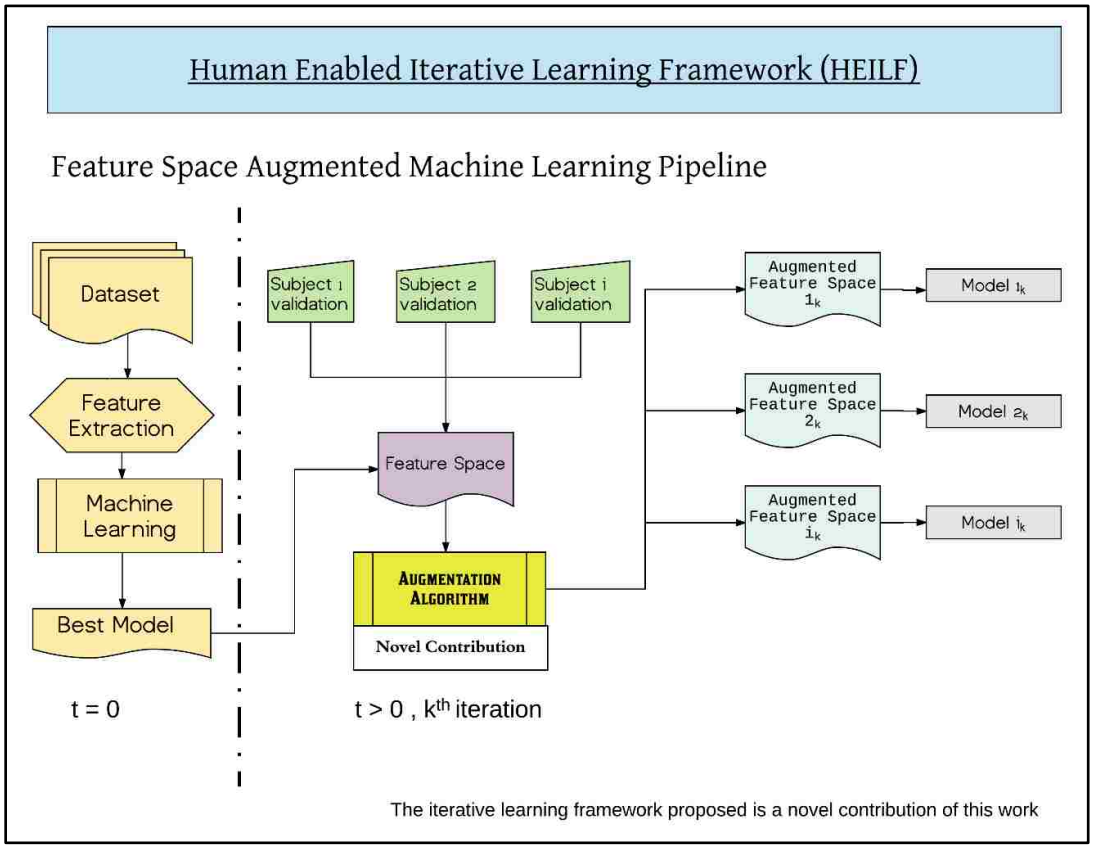


Figure 33: Human Enabled Iterative Learning Framework

3.9 Evaluation

The machine learning classifier to predict dominant emotional state is validated on the test data using 5-fold cross validation, 10-fold cross validation, 15-fold cross validation, 20-fold cross validation, 25-percent holdout validation and 50-perent holdout validation. The 50% holdout validation yields the most conservative estimates. The results from 50-percent holdout validation are presented in this work. The Levenberg-Marquart [52] algorithm used in the neural net based regression uses 15% of the data as the test set and 15% as a validation set. The evaluation parameters for the regression are presented in 3.7 .

3.9.1 Six-Category Classification: kNN

Table 8: Hyper parameter optimization – minimize 5-fold cross validation loss- 6 categories

Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	NumNeighbors	Distance
1	Best	0.2216	1.7734	0.2216	0.2216	2	seuclidean
2	Accept	0.22346	0.23259	0.2216	0.22253	32	seuclidean
3	Accept	0.8175	0.17145	0.2216	0.22255	5	hamming
4	Accept	0.83426	0.14162	0.2216	0.22265	45	hamming
5	Accept	0.83426	0.36701	0.2216	0.22271	1	spearman
6	Accept	0.25326	0.11975	0.2216	0.22266	1	minkowski
7	Accept	0.25512	0.17064	0.2216	0.22264	1	mahalanobis
8	Accept	0.24767	0.11067	0.2216	0.22263	4	chebychev
9	Best	0.21601	0.10666	0.21601	0.21621	7	cityblock
10	Accept	0.49721	0.10711	0.21601	0.21624	4	correlation
11	Accept	0.66294	0.111	0.21601	0.21628	265	cosine
12	Accept	0.64432	0.098836	0.21601	0.21631	1	jaccard
13	Accept	0.25326	0.105	0.21601	0.2163	1	euclidean
14	Accept	0.68715	0.11541	0.21601	0.21607	269	cityblock
15	Accept	0.2514	0.10574	0.21601	0.21609	1	cityblock
16	Accept	0.47672	0.11599	0.21601	0.21608	89	chebychev
17	Accept	0.24953	0.11552	0.21601	0.21609	20	minkowski
18	Accept	0.24209	0.10358	0.21601	0.2161	25	euclidean
19	Accept	0.43575	0.10714	0.21601	0.2161	32	mahalanobis
20	Accept	0.2514	0.10196	0.21601	0.21611	1	chebychev
Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	NumNeighbors	Distance
21	Best	0.19553	0.10074	0.19553	0.19558	8	seuclidean
22	Accept	0.68901	0.10931	0.19553	0.19554	266	euclidean
23	Accept	0.23091	0.10438	0.19553	0.19554	6	euclidean
24	Accept	0.69088	0.11483	0.19553	0.19553	269	minkowski
25	Accept	0.22533	0.11147	0.19553	0.19554	5	minkowski
26	Accept	0.68156	0.10961	0.19553	0.19549	268	seuclidean
27	Accept	0.2365	0.10215	0.19553	0.19548	3	cityblock
28	Accept	0.20857	0.12354	0.19553	0.19707	16	seuclidean
29	Accept	0.73371	0.10745	0.19553	0.19715	268	correlation
30	Accept	0.83426	0.1072	0.19553	0.19739	269	jaccard

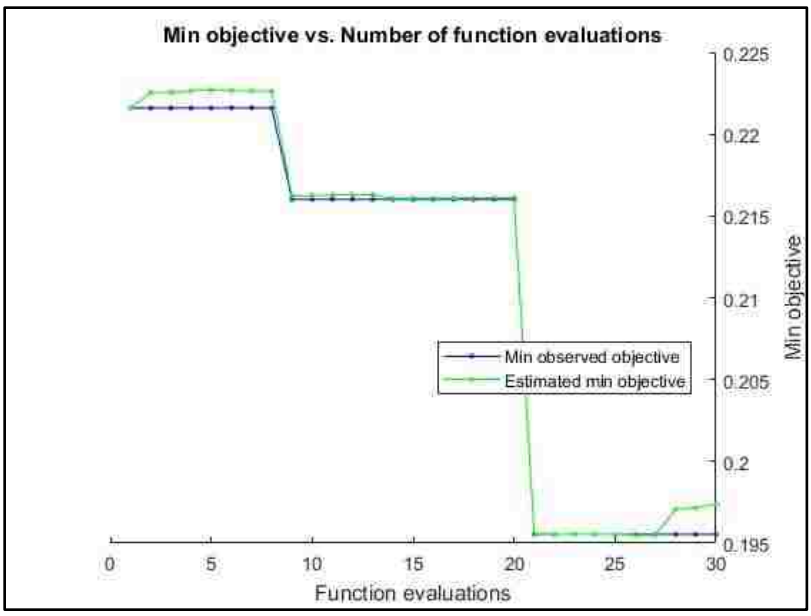


Figure 34: Hyperparameter optimization- cross-validation loss, 6-Categories

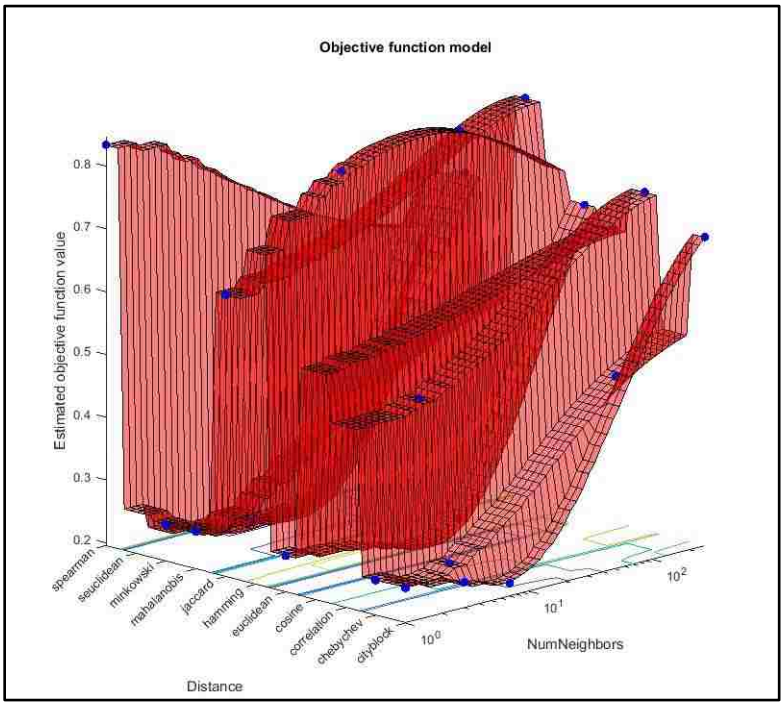


Figure 35: Hyperparameter optimization- Objective function model, 6-Categories

3.9.2 Three-Category Classification; kNN

Table 9: Table 8: Hyper parameter optimization – minimize 5-fold cross validation loss- 3 categories

Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	NumNeighbors	Distance
1	Best	0.11373	0.17709	0.11373	0.11373	2	seuclidean
2	Accept	0.16863	0.10947	0.11373	0.11731	20	seuclidean
3	Accept	0.52157	0.1052	0.11373	0.13818	4	hamming
4	Accept	0.66667	0.10337	0.11373	0.15446	27	hamming
5	Best	0.10196	0.098183	0.10196	0.10201	1	seuclidean
6	Accept	0.66667	0.15905	0.10196	0.10209	1	spearman
7	Accept	0.12157	0.09919	0.10196	0.10203	1	minkowski
8	Accept	0.53333	0.10048	0.10196	0.10206	128	minkowski
9	Accept	0.11765	0.10105	0.10196	0.10204	1	mahalanobis
10	Accept	0.53725	0.098461	0.10196	0.10204	119	mahalanobis
11	Accept	0.1451	0.1097	0.10196	0.10202	6	cityblock
12	Accept	0.54118	0.099774	0.10196	0.10201	126	cityblock
13	Accept	0.11765	0.12709	0.10196	0.10202	1	cityblock
14	Accept	0.12941	0.099458	0.10196	0.10201	1	chebychev
15	Accept	0.18039	0.11134	0.10196	0.10201	15	chebychev
16	Accept	0.29804	0.092394	0.10196	0.10201	2	correlation
17	Accept	0.16863	0.092516	0.10196	0.102	2	cosine
18	Accept	0.4	0.094657	0.10196	0.102	57	cosine
19	Accept	0.12157	0.11794	0.10196	0.102	1	euclidean
20	Accept	0.16863	0.096956	0.10196	0.102	11	euclidean
Iter	Eval result	Objective	Objective runtime	BestSoFar (observed)	BestSoFar (estim.)	NumNeighbors	Distance
21	Accept	0.35294	0.10678	0.10196	0.102	1	jaccard
22	Accept	0.56471	0.09668	0.10196	0.102	127	correlation
23	Accept	0.53333	0.099576	0.10196	0.102	128	euclidean
24	Accept	0.15294	0.097213	0.10196	0.102	3	euclidean
25	Accept	0.16078	0.096731	0.10196	0.102	3	chebychev
26	Accept	0.5451	0.10019	0.10196	0.10199	127	chebychev
27	Accept	0.15294	0.096013	0.10196	0.10428	2	cityblock
28	Accept	0.16863	0.095133	0.10196	0.10495	3	mahalanobis
29	Accept	0.53333	0.09844	0.10196	0.10295	128	seuclidean
30	Accept	0.15294	0.095923	0.10196	0.10198	7	seuclidean

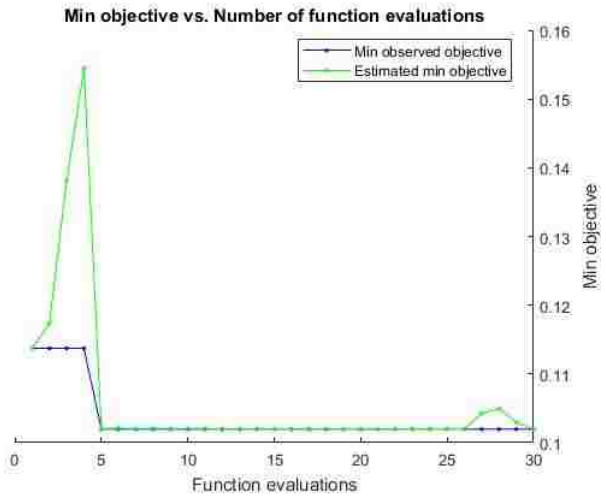


Figure 36: Figure 33: Hyperparameter optimization- cross-validation loss, 3-Categories

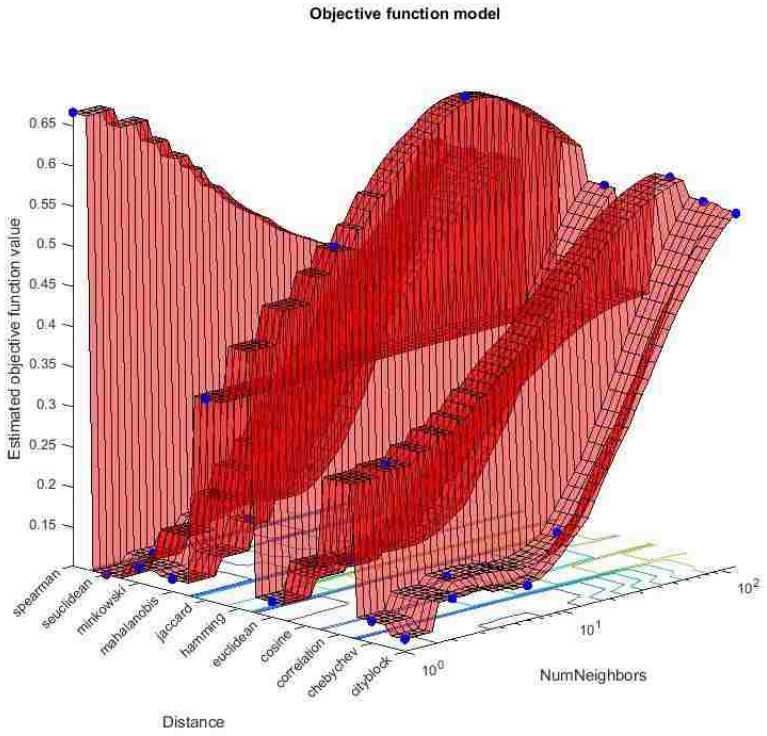


Figure 37: Hyperparameter optimization- Objective function model, 3-Categories

3.9.3 Comparing our work to other state of the art

Table 10: Comparing results to other significant works

Taxonomy	Research	Evaluation Parameters				
		Accuracy	Computational Effeciency	Prediction Granularity	Realtime Implementation	Training Data
Wearable Sensors	i-Feel (our system) with an ITERATIVE LEARNING ALGORITHM & EMOTION SPECTRUM	92% (3 category) 87% (6 category)	Classifier on 3 dimensional space	3/6 categories 1x7 emotional spectrum	Sutible	85 subjects. 9 signals per subject 9 surveys per subject Video stimulus
	Emotion Recognition Using Physiological Signals: Laboratory vs. Wearable Sensors Ragot et al. 2017 [similar accuracy between soa. lab sensors and E4]	66% valance 70% arousal	Classifier on 9 dimensional space	9 categories valance-arousal	Sutible	19 subjects. 45 signals per subject Image stimulus (IAPS)
	An Emotion Recognition System Based on Physiological Signals Obtained by Wearable Sensors . He et al. 2016	57.34% valance	Classifier on 145 dimensional space	4 categories	not-sutible	11 subjects
	Emotion recognition through physiological signals for human-machine communication. Maaoui et al. 2010	92%	Classifier on 30 dimensional space	6 categories	not-sutible	10 subjects 6 signals per subject Image stimulus 1 image/ 5 seconds
	From the lab to the real-world: An investigation on the influence of human movement on Emotion Recognition using physiological signals .Xu et al. 2017	91.25	N/A	5 categories valance-arousal	not -sutible	8 subjects Image stimulus (IAPS) 5 | 30 sec signals/ subject
	Toward machine emotional intelligence: Analysis of affective physiological state.Picard et al. 2001	65.3 % (5 category) 88.3 % (3 category)	Classifier on 24 dimensional space	5/3 categories	not-sutible	1 subject

4.0 PROBLEM 2

Problem Statement: Combining Image features and text features from optical character recognition (OCR) to create a hybrid classifier for robust image prediction in an iterative machine learning framework. The goal of this study is to improve the accuracy of existing image recognition by leveraging text features from the images. As humans, we perceive objects using colors, dimensions, geometry and any textual information we can gather. Current image recognition algorithms rely exclusively on the first 3 and do not use the textual information. This study develops and tests an approach that allows for inclusion on the text features in the learning algorithm. The study includes an iterative learning layer that allows for the system to improve over time through human machine interaction.

Data

The data set used for this work is the Asset and Tag images dataset curated from the industry sponsor for this project. The data set contains about 200,000 images from building assets. Building assets include industrial equipment such as HVAC units, PTACKS, Microwaves etc. There are a total of 15 classification categories that make up over 92 percent of the assets. These 15 categories constitute the image labels. Each asset has 2 images, image1 being an isometric view of the asset and Image 2 being a close-up of a Tag with Manufacturer name and other model details. There is a tremendous variability (within the same class) within the asset images which can be attributed to image quality (illumination, scale, and perspective), age of equipment, and variations due to multiple manufacturers and models.

4.1 Related Works and Taxonomy

Convolutional Neural Nets and Image recognition

Convolutional neural networks can be traced back to Hopfield et al. in 1982[53]. However, the foundation of all CNN's can be traced back to the back-propagation algorithm proposed in 1986 by Rumelhart et al[54]. The first practical application was published in 1998 by LeCun et al. where the neural net LeNet 5 was used to classify the MINST dataset[55], [56].The work lead to a 99.2 % accuracy. Hubel et al. proposed the architecture of human visual perception[57] . The paper defines visual perception mechanism as a layered architecture of neurons within the human brain. This inspired scientists to reconstruct this architecture to aid computer vison.

Input Data: CNN's are used on images in computer vison. A 3-channel image (RGB) contains 3 matrices representing RGB intensities of nxm pixels. Given we have 8-bit pixels, each pixel represents a value between 0-255.

Convolution Kernels: A convolution kernel (also called a filter) is a matrix of real valued entries that operates on the entire image, transforming the information contained in the pixels to information used for analysis. The convolution of the kernel with the image yields activation maps. These are the regions where the features specific to the kernel are detected. The values contained in the convolution kernel iterate over the training set leading to a kernel that best identifies regions of the image suitable for feature extraction.

Kernel Operations

1. The kernel of size mxn is convolved with image patches of the same dimension.
2. The convolved entry (real number) becomes an entry in the activation matrix. The value is normalized by dividing by the dimension of the convolution kernel. (the convolution value is obtained by the dot product of the image patch and the kernel)
3. The kernel is then convolved with another mxn patch by sliding it over the patch by a stride value (number of columns), till the activation matrix for the entire image is complete.

Convolution Layer: The act of convolving an image with many filters and creating a stack of featured images is called a convolution layer. It's called a "layer" since it's an operation, that can be stacked with other layers.

Pooling (shrinking the image stack): Pooling involves picking a window size and a stride length. The window is walked across the filtered image and the maximum value is recorded for each window.

Rectified Linear Unit: RELU is a normalization operation. Every negative value is changed to zero.

Deep stacking: The convolution, ReLU and Pooling layers are stacked man times, leading to a filtered (significant dimension reduction) of the original image.

Fully Connected Layer: the stacked filtered images are converted to a list (1 dimensional) with each vector having a target label. Fully connected layers may also be stacked.

Backpropagation and Gradient descent: Each feature pixel (convolution layer) and voting weight (fully connected layer) are adjusted based on the error. The lowest point on the error gradient curve is used to assign the weights.

Hyperparameters (user defined parameters): User inputs, these include,

- Number of features (kernels)
- Size of features
- Pooling window size
- Pooling window stride
- Number of neurons in the fully connected layer
- Number and order of layers

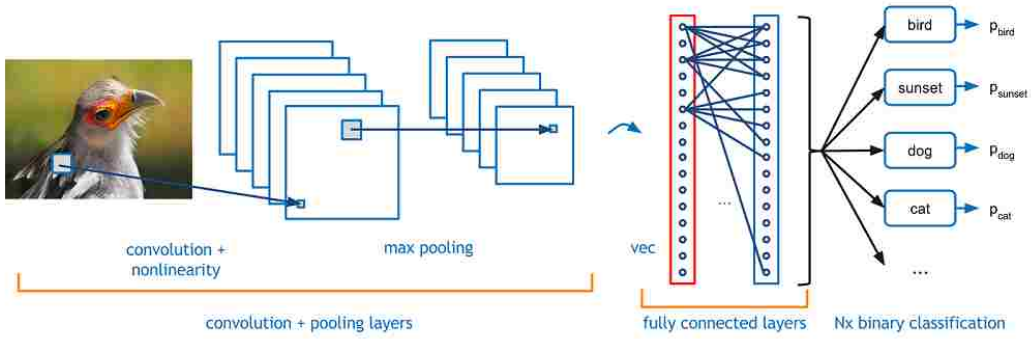


Figure 38 : Basic CNN Architecture [58]

There has been tremendous research in the field of CNN's. ImagNet is one of the largest opensource image database[59]. The database currently contains over 14 million images from 1000 categories. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an

yearly competition that features object localization for thousand categories, object detection for two hundred categories and object localization for 30 fully labelled categories[59]. MatConvNet is a Matlab toolbox for implementing the state of the art CNN's in Matlab[60]. The results from the most popular models are presented in the table below. The prediction is made using a CNN that leads to a multinomial distribution of the predicted classes. The top-1 score checks if the target class is the same as the class with the highest probability. The top-5 score checks if the target class is the same as the top 5 predicted classes.

Table 11: Model Performance on ILSVRC 2012 validation data[60]

model	introduced	top-1 err.	top-5 err.	images/s
matconvnet-vgg-verydeep-16	2014	28.3	9.5	200.9
vgg-verydeep-19	2014	28.7	9.9	166.2
vgg-verydeep-16	2014	28.5	9.9	200.2
googlenet-dag	2014	34.2	12.9	770.6
matconvnet-alex	2012	41.8	19.2	2133.3

Transfer Learning: Transfer learning is the process of taking a pretrained CNN and finetuning it for another dataset. In this context, the pretrained CNN can be thought of as a feature extractor. The layer preceding the fully connected layer is used as the feature for any given image (image decomposed into an array). Thus, a set of labelled images can be encoded into the CNN feature space and machine learning algorithms can be used to train classifiers on this feature space. Training a CNN from scratch is an extremely compute intensive process. This can be attributed to the iterative nature of training that employs back propagation and gradient descent to generate the convolution kernel. Feature extraction using transfer learning takes advantage of the well-developed CNN architecture and is significantly less compute intensive. In [61], [62] the authors show that the features extracted from the activation of a deep convolutional network can be trained in a fully supervised learning environment and can be repurposed for novel tasks.

AlexNet: In [63] the authors trained a deep convolutional neural net to classify 1.2 million high-resolution images from the ImageNet LSVRC 2010[59] into 1000 categories. The neural net contains 60 million parameters and 650,000 neurons is composed of five convolutional layers with intermediate pooling layers and three fully connected layers. The network won under the top-5 test error rate category at the ILSVRC 2012 [59]. This was one of the foundations of GPU trained CNN's.

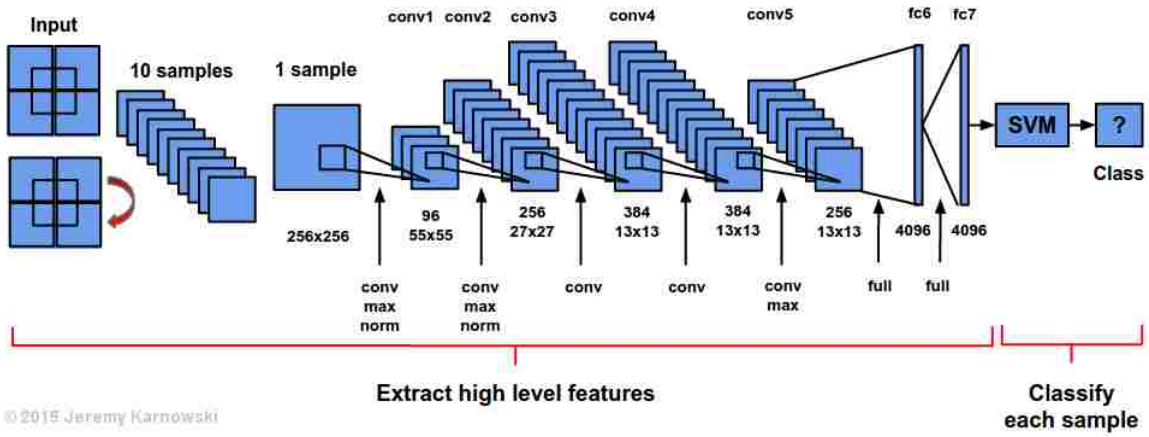


Figure 39:AlexNet Transfer Learning Framework[64]

VGG-VD-19: In [65] the authors investigate the effect of the depth of a CNN on the accuracy in large scale image recognition problems. The authors show that significant increase in the prediction accuracy can be achieved using weight layers with a depth of 16-19 layers. The mode placed first in the ILSVRC 2014 in the localization and classification challenges[59]. The network configuration is presented in the table below.

Table 12: CovNet Configuration VGG-VD [65]

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

The fc7 layer from AlexNet[63] and the FC-1000 layer from VGG-VD-19[65] are used as feature extraction layers. fc7 encodes an image into a 1x4096 feature vector and FC-1000 encodes an image as a 1x1000 vector.

Image Classification using Visual Bag of words

In [66] the authors describe a method of encoding images into features using a visual bag of words model. The algorithm generates a histogram of visual word occurrences that the image is composed of. The steps outlined in the workshop[66] are,

1. Separate the images into a test and training set
2. Create a Bag of Features: The bag of features is creating a vocabulary of visual words using k-means clustering. The vocabulary is generated by using feature descriptors extracted from the training set. The k-means algorithm groups the descriptors into user defined clusters. The feature detectors used are SIFT(scale invariant feature transform)[67], [68] and SURF(speeded up robust features)[69].
3. Each image is encoded into a feature vector (1xnumber of clusters) based on the occurrence (frequency) of visual words within the image.
4. A machine learning classifier is used to train on the encoded image space.
5. The classifier is validated using the model with the test set.

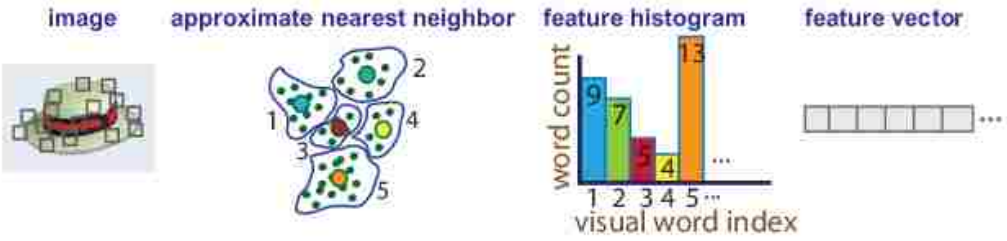


Figure 40: Encoding Images into bag of visual words[70]

4.2 Research Objectives

Encode an image based on a text based feature space and train a classifier on re-encoded images and finally compare its performance (accuracy and computational efficiency) to the state of the art classification algorithms,

1. Convolutional Neural Nets (transfer learning and feature extraction)
    i. Alexnet
    ii. VGG-VD19
2. Key point Detection
    iii. SIFT
    iv. SURF
3. Create a hybrid feature space combining image and text features and evaluate its performance relative to CNN's and key point detection based methods. This is accomplished by,
    i. Creating a vocabulary for the problem space
    ii. Converting the vocabulary to a document term matrix
    iii. Re-encoding images into a lower dimension (15) space
        - Creating a boosting algorithm
    iv. Training a classifier on the new space and a combination of the new space and the image features from CNN.

4.3 Methodology

In this section, we talk about the hardware used, the general process of encoding images to a text based feature space and the MSER (Maximally stable extermal images) for text extraction from natural scenes.

4.3.1 Hardware Used

For this work a 2 PC's with dual hex core Xeon processors (48 cores with hyper threading), 512 gigabytes of ram and an NVidia GeForce 1080Ti (3500 CUDA cores) were used. The machines were a part of a Matlab Distributed Computing Server.

The preprocessing and natural image OCR and the machine learning algorithms were completed using the CPU cores while the Convolutional Deep Learning Nets were run on the Nvidia GPU using existing Matlab compatible CUDA libraries.

4.3.2 Encoding an Image onto a text based feature space

Encoding an image into a text based feature has many implications for the classification problem,

1. It allows for significant dimension reduction compared to other algorithms. The image of a specific size is encoded into a 1x1000 vector in state of the art Convolutional Neural Nets such as VGG-VG19 and Alex Net at the final fully connected layer. The layers before that have much higher dimensions (e.g. fc7 in AlexNet encodes an image into a vector with over 4000 dimensions). Text based encoding encodes an image of any size to a 1xm vector (m is the number of classification categories)

2. In this specific problem (15 classification categories) the machine learning algorithm is trained on a 15-dimensional feature space. This allows significantly faster training and re-training times (discussed in the Results section) framework.

4.3.3 MSER Algorithm (maximally stable extermal regions)

MSER is a blob detection algorithm proposed by Matas et al[71]. The algorithm extracts co-variant regions from an image. The motivation behind MSER is based on identifying regions that show minimal variation across a wide range of thresholds. All pixels below a threshold are white and the ones above are black. The set of connected components across the threshold are the sets of extermal regions detected.

External region implies that all pixels within the boundary have a lower or higher intensity compared to the pixels outside the region boundary. The steps outlined in [72] are,

1. Simple luminance thresholding of the image sweeping the threshold intensity from back to white.
2. Extraction of Extermal Regions
3. Iteratively find the threshold at which the region is maximally stable.
4. Keep the region descriptors as features

The hyperparameters (user defined) include maximum area, minimum area and maximum variation of pixel intensity within the region.

Image I is a mapping $I : \mathcal{D} \subset \mathbb{Z}^2 \rightarrow \mathcal{S}$. Extremal regions are well defined on images if:

1. $\mathcal{S}$ is totally ordered, i.e. reflexive, antisymmetric and transitive binary relation $\leq$ exists. In this paper only $S = \{0, 1, \ldots, 255\}$ is considered, but extremal regions can be defined on e.g. real-valued images ($\mathcal{S} = R$).
2. An adjacency (neighbourhood) relation $A \subset \mathcal{D} \times \mathcal{D}$ is defined. In this paper 4-neighbourhoods are used, i.e. $p, q \in \mathcal{D}$ are adjacent (pAq) iff $\sum_{i=1}^{d} |p_i - q_i| \leq 1$.

Region $\mathcal{Q}$ is a contiguous subset of $\mathcal{D}$, i.e. for each $p, q \in \mathcal{Q}$ there is a sequence $p, a_1, a_2, \ldots, a_n, q$ and $pAa_1, a_iAa_{i+1}, a_nAq$.

(Outer) Region Boundary $\partial\mathcal{Q} = \{q \in \mathcal{D} \setminus \mathcal{Q} : \exists p \in \mathcal{Q} : qAp\}$, i.e. the boundary $\partial\mathcal{Q}$ of $\mathcal{Q}$ is the set of pixels being adjacent to at least one pixel of $\mathcal{Q}$ but not belonging to $\mathcal{Q}$.

Extremal Region $\mathcal{Q} \subset D$ is a region such that for all $p \in \mathcal{Q}, q \in \partial\mathcal{Q} : I(p) > I(q)$ (maximum intensity region) or $I(p) < I(q)$ (minimum intensity region).

Maximally Stable Extremal Region (MSER). Let $\mathcal{Q}_1, \ldots, \mathcal{Q}_{i-1}, \mathcal{Q}_i, \ldots$ be a sequence of nested extremal regions, i.e. $\mathcal{Q}_i \subset \mathcal{Q}_{i+1}$. Extremal region $\mathcal{Q}_{i^*}$ is maximally stable iff $q(i) = |\mathcal{Q}_{i+\Delta} \setminus \mathcal{Q}_{i-\Delta}|/|\mathcal{Q}_i|$ has a local minimum at i^* ($|.|$ denotes cardinality). $\Delta \in \mathcal{S}$ is a parameter of the method.

Figure 41: Algorithm Description[71]

4.4 Algorithm Pipeline

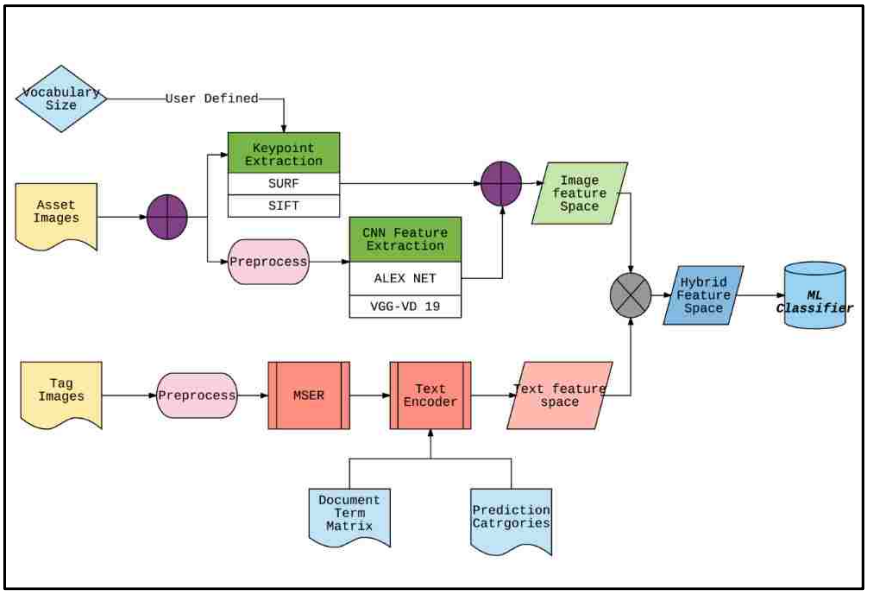


Figure 42: Algorithm pipeline for hybrid feature space

4.4.1 Text extraction using MSER and Natural Image OCR -Illustrated Example

MSER algorithm is extremely effective in detecting text in unstructured images. An unstructured image contains random scenarios[73]. Bill boards are a common example of unstructured images since they have a combination of images and text. Traditional OCR performs well with text documents but poorly with unstructured images. The tag images from the industrial equipment image database is a good candidate for MSER application.

Figure 43: Sample Asset Image

Detecting extermal regions: The MSER feature detector is used to identify extermal regions. The image is converted to grayscale and a threshold for image sweeping is defined. This step detects the possible candidates for extermal regions.

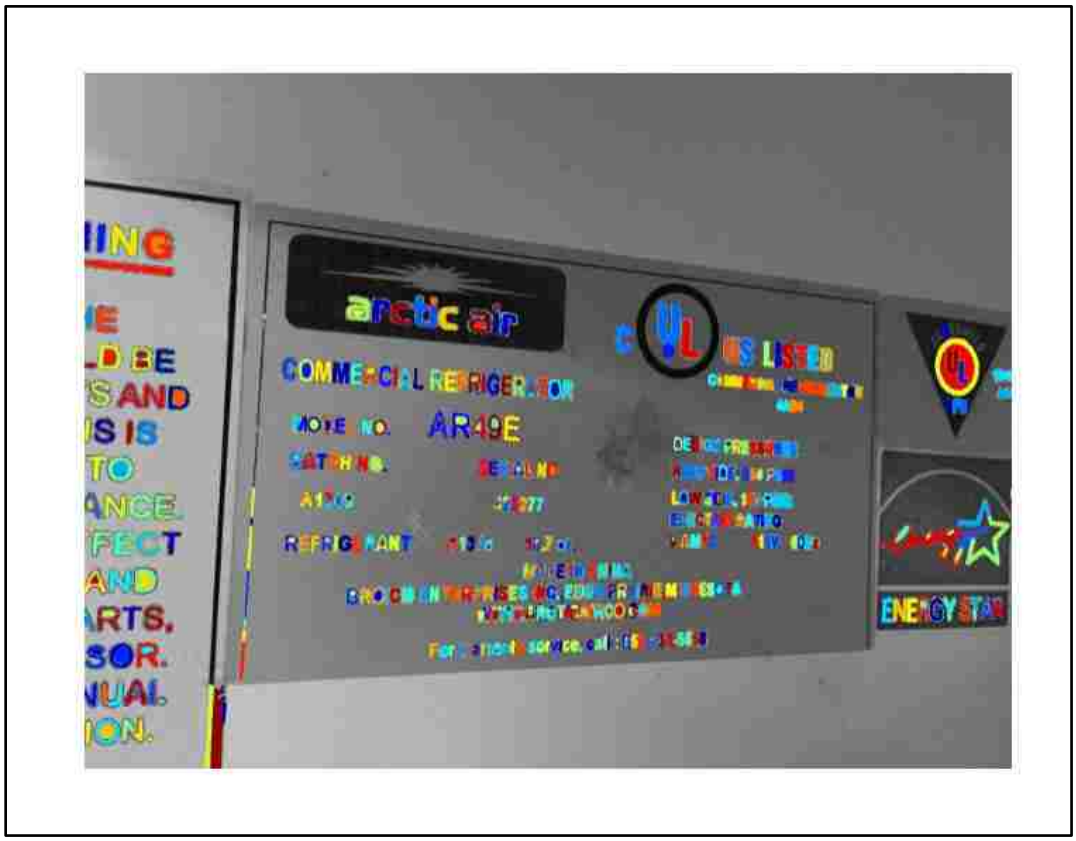


Figure 44: Detected MSER regions

Removing Non-Text Regions: Since the image might contain non-text MSER regions geometric properties can be used to remove non-text regions. This is accomplished by a rule based approach combined with a machine learning classifier that distinguished between text and non-text regions based on region properties[74]. Authors in [75], [76] present geometric properties that can distinguish between next and non-text regions detected by MSER

algorithm. These region properties as defined in [77] include, aspect ratio, eccentricity (Returns a scalar that specifies the eccentricity of the ellipse that has the same second-moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1. (0 and 1 are degenerate cases. An ellipse whose eccentricity is 0 is a circle, while an ellipse whose eccentricity is 1 is a line segment.), Euler number (Returns a scalar that specifies the number of objects in the region minus the number of holes in those objects.), Extent (Returns a scalar that specifies the ratio of pixels in the region to pixels in the total bounding box. Computed as the Area divided by the area of the bounding box) and Solidity (Returns a scalar specifying the proportion of the pixels in the convex hull that are also in the region. Computed as Area/Convex Area). Stroke width variation is also used as a metric to identify text regions based on the approach proposed in [76]. We also added a rule based approach that removes regions with number of pixels less than the median pixels per regions. This filter significantly improved the detected text for the tag image dataset.

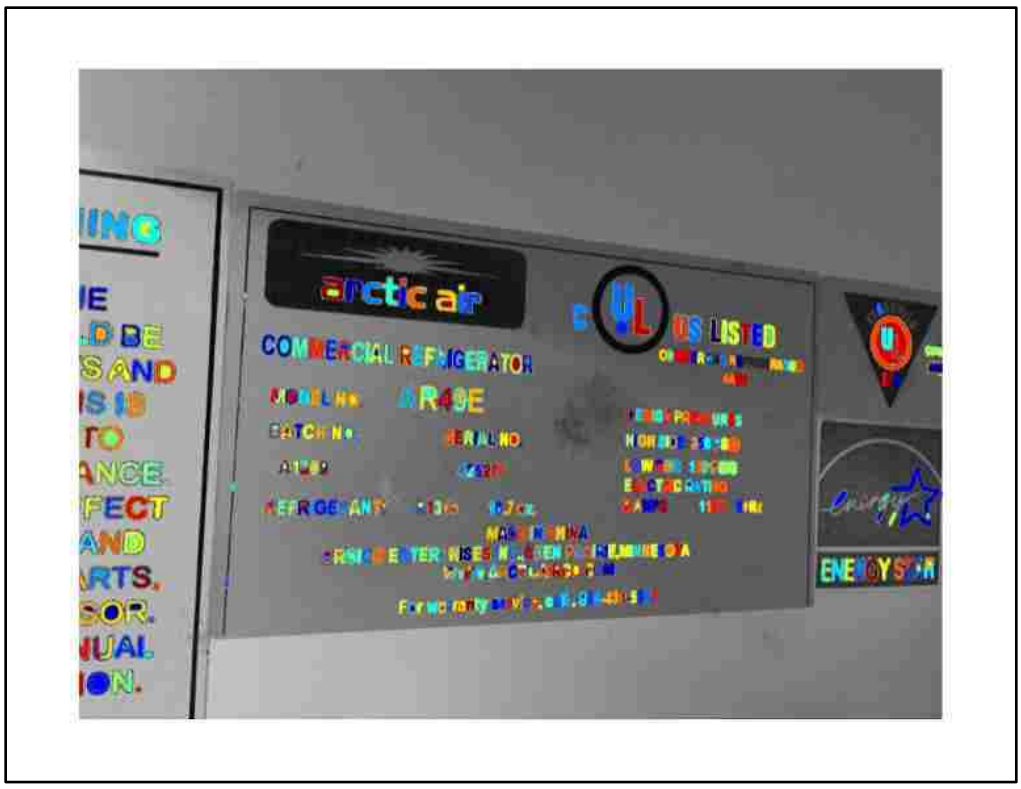


Figure 45: Filtered text based regions

Merging text regions : The MSER regions detected contain individual characters. The goal of OCR is to detect complete words and sentences that can be used to gather context about the text. This is accomplished by creating and iteratiely expanding boingboxes to detect overlap.

Figure 46: Creating bounding box around detected regions

The bounding boxes are expanded by a small amount in the x direction (since the words/ sentences go left to right). The bounding boxes are then collapsed based on a user defined overlap ratio. The detailed code can be found in the Appendix. The OCR (pretrained classifier) is run on individual bounding boxes to predict words and the probability of prediction.

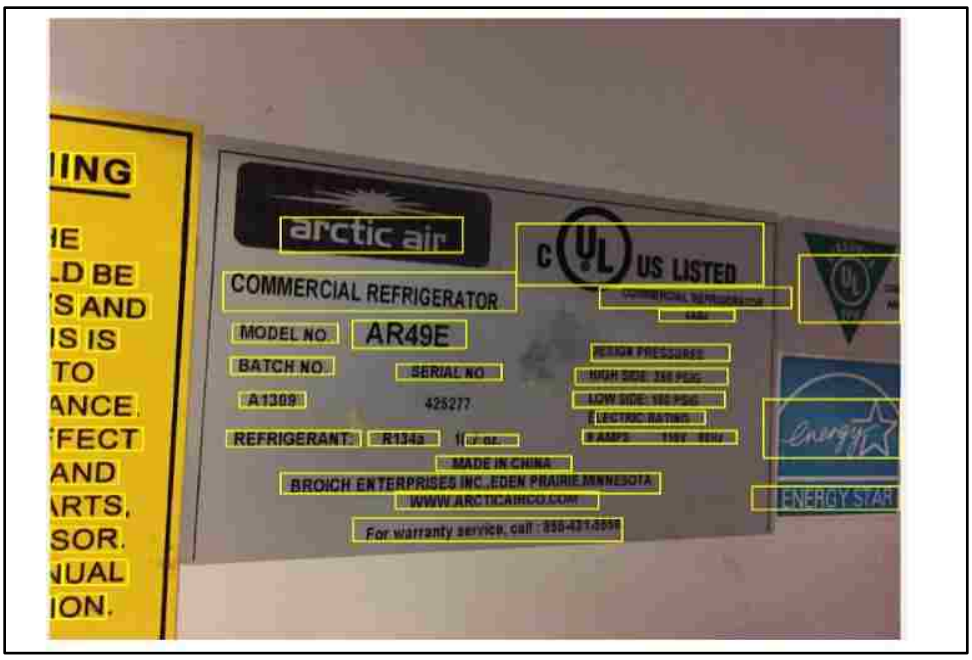


Figure 47: Expanded Bounding Box

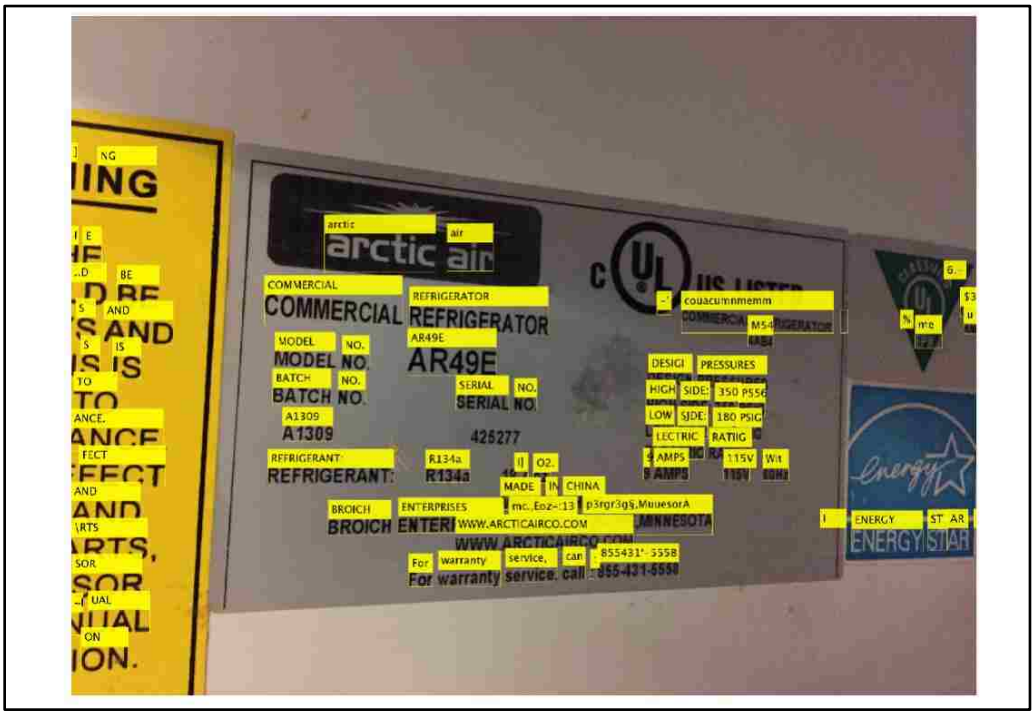


Figure 48: Detected text

4.4.2 Data Cleaning and Preprocessing

The text extracted from each image is saved as a table with terms, bounding box coordinates and prediction confidence as columns. The preprocessing steps include lemmatization, removal of stop words, punctuations and special characters. The tables from each category are then concatenated creating 15 tables, 1 for each classification category. As an added preprocessing measure, the terms with frequency less than the mean term frequency for each classification category are removed. The primary reason for this step is to create a vocabulary (representing each category) of manageable size. This decreases the number of terms by 60 percent which decreases the sparsity of the document term matrix significantly. The document term matrix is read into the memory as a broadcast variable in an SPMD (single program multiple data) framework, i.e. it cannot be dumped until the batch process is completed. The size of the document term matrix has memory implications, especially in GPU computing where the GPU memory is a performance bottle neck.

4.4.3 Document Term Matrix

The concatenated term tables for each category are converted to a corpus of 15 documents, each representative of a classification category. A document term matrix is a numeric matrix the categories as rows and terms as columns. For this study, we generate a 15 x 150000 matrix. An element i, j represents the frequency of term j for the i^{th} document (classification category). Despite the preprocessing measures, the DTM is a sparse matrix. The document term matrix is used to encode images (matrices of any size) into a text based feature space. The columns of the DTM are marginalized (column sum equals 1). Each column (term) now represents the probability of a term occurring in a specific corpus. While most of, much of

terms in the DTM are do not have dictionary accuracy, the goal is to capture word permutations that might be representative of a category. For instance, a natural image with the word “refrigeration” might yield incomplete terms such as “referi”, “ref”, “referige” etc. where each term represents a column in the DTM, with a large enough data set (such as TELS) it is possible to create a granular feature space, i.e. each permutation appears with high frequency in the DTM that makes the re encoding effective. Moreover, it is possible to boost the results by combining the contributions of each permutation to the feature space.

4.4.4 Encoding Images onto a text based feature space

Feature encoding is a form of dimension reduction technique. In this work, we encode an mxn matrix into a 1x15 vector using the document term matrix. Each element of this vector represents the volume of text in the image that comes from a certain classification category (15 categories). The encoded images create a 15-dimension feature space that is used to train machine learning classifiers.

Table 13: Document-term matrix

	T1	T2	T3	T4	T5	T m
Cat1	.5	.3	.1	….	……	……..
Cat2	.2	.7	.2	…….	……..	………..
Cat3	.3	0	.7	……..	……..	……..
.	……..	……..	……..	……..	……..	……..
.	……..	……..	……..	……..	……..	……..
Cat 15	……..	……..	……..	……..	……..	……..

The rows of Table 1.0 represent the 15 classification categories while the columns represent the all terms (from the vocabulary of extracted text). A new image undergoes the Natural Image OCR aided by the MSER algorithm. The text from the image is converted to a term frequency matrix.

Table 14: Term frequency table for sample image

Term	Frequency of the m^{th} term
T_1	f_1
T_2	f_2
T_m	f_m

The sample image has m terms $\mathbf{T_1}$ to $\mathbf{T_m}$ with frequencies $\mathbf{f_1}$ to $\mathbf{f_m}$. Let V define a n dimensional feature space on to which the image is encoded.

$$V = [V_1 \ V_2 \ V_3 \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots V_n]$$

Each element of this vector corresponds to, the volume of the image text that belongs to a specific category. E.g. the second term of V would represent the volume of image text that from Category 2. Hence, if n is the number of categories and m is the total terms in the DTM, A. Then,

$$V_1 = f_1 A_{11} + f_2 A_{12} + f_3 A_{13} + \ldots\ldots\ldots f_m A_{1m.}$$

More generally,

For i = 1.....n

V = length (n)

4.4.5 NIOCR-Function

The MESR text detection and the image classification algorithm described in 4.4.4 were implemented in a single function for implementation in an SPMD framework.

```
function [ vec ] = NIocr( img,dtm,vocabulary)% imgloc - image location
try

vec =[];
colorImage = imread(img);
I = rgb2gray(colorImage);

%% Detect MSER regions.
[mserRegions, mserConnComp] = detectMSERFeatures(I, ...
    'RegionAreaRange',[30 14000],'ThresholdDelta',.8,'MaxAreaVariation',0.1);

%% Use regionprops to measure MSER properties
mserStats = regionprops(mserConnComp, 'BoundingBox', 'Eccentricity', ...
    'Solidity', 'Extent', 'Euler', 'Image');
if isempty(mserStats)~=1

%% Compute the aspect ratio using bounding box data.
bbox = vertcat(mserStats.BoundingBox);
w = bbox(:,3);
h = bbox(:,4);
aspectRatio = w./h;

%% Threshold the data to determine which regions to remove. These thresholds
% may need to be tuned for other images.
filterIdx = aspectRatio' > 2;
filterIdx = filterIdx | [mserStats.Eccentricity] > .99 ;
filterIdx = filterIdx | [mserStats.Solidity] < .1;
%% Remove regions
mserStats(filterIdx) = [];
mserRegions(filterIdx) = [];
%% Bounding Boxes
%Get bounding boxes for all the regions
bboxes = vertcat(mserStats.BoundingBox);

%% Added by PRS : Remove non-informative blocks
%non-informative blocks-the blocks with pixels less than the median pixels.

for j = 1:numel(mserStats)
  [mserStats(j).pixels]= numel(mserStats(j).Image);
end
if isempty(mserStats) ~= 1
med  = median((cat(1.,mserStats.pixels)));
k    =  find(cat(1.,mserStats.pixels)<1*med);
mserStats([k])=[];
bboxes = vertcat(mserStats.BoundingBox);
```

```
%% Bounding Box: Convert from the [x y width height] bounding box format to
the [xmin ymin]
%xmax ymax] format for convenience.
xmin = bboxes(:,1);
ymin = bboxes(:,2);
xmax = xmin + bboxes(:,3) - 1;
ymax = ymin + bboxes(:,4) - 1;
bboxes= [xmin ymin xmax-xmin+1 ymax-ymin+1];

%% Bounding Box: Expand the bounding boxes by a small amount.
expansionAmount = 0.04;
xmin = (1-expansionAmount) * xmin;
%ymin = (1-expansionAmount) * ymin;
xmax = (1+expansionAmount) * xmax;
%ymax = (1+expansionAmount) * ymax;
% Clip the bounding boxes to be within the image bounds
xmin = max(xmin, 1);
ymin = max(ymin, 1);
xmax = min(xmax, size(I,2));
ymax = min(ymax, size(I,1));
%Show the expanded bounding boxes
expandedBBoxes = [xmin ymin xmax-xmin+1 ymax-ymin+1];
%IExpandedBBoxes =
insertShape(colorImage,'Rectangle',expandedBBoxes,'LineWidth',3);
%figure
%imshow(IExpandedBBoxes)
%title('Expanded Bounding Boxes Text')

%% Bounding Box: Compute the overlap ratio | Merge boxes
overlapRatio = bboxOverlapRatio(expandedBBoxes, expandedBBoxes);
% Set the overlap ratio between a bounding box and itself to zero to
% simplify the graph representation.
n = size(overlapRatio,1);
overlapRatio(1:n+1:n^2) = 0;
%% Create the graph
g = graph(overlapRatio);
%plot(g);

%% Find the connected text regions within the graph
componentIndices = conncomp(g);
% Merge the boxes based on the minimum and maximum dimensions.
xmin = accumarray(componentIndices', xmin, [], @min);
ymin = accumarray(componentIndices', ymin, [], @min);
xmax = accumarray(componentIndices', xmax, [], @max);
ymax = accumarray(componentIndices', ymax, [], @max);

%% Compose the merged bounding boxes using the [x y width height] format.
textBBoxes = [xmin ymin xmax-xmin+1 ymax-ymin+1];

%% Bounding Box: Remove bounding boxes that only contain one text region
numRegionsInGroup = histcounts(componentIndices);
textBBoxes(numRegionsInGroup == 1, :) = [];
```

```
%% Preprocess the image to fine-tune results
ocrtxt = ocr(I, textBBoxes,'TextLayout','Block');

if isempty(ocrtxt)~= 1

    %% Create a Table
coordinates = vertcat(ocrtxt.WordBoundingBoxes); %box coordinates
words = (vertcat(ocrtxt.Words));
confidence = num2cell(vertcat(ocrtxt.WordConfidences));
xmin = num2cell(coordinates(:,1));
ymin = num2cell(coordinates(:,2));
xmax = num2cell(coordinates(:,3));
ymax = num2cell(coordinates(:,4));
table = horzcat(words,confidence,xmin,ymin,xmax,ymax);

tmp =table(:,1);
t1 = tabulate(tmp); % COMMENT| NOTE : The 3rd column - percent can also be used
for re-encoding purposes
t1 = cell2table(t1);
t1.colind = zeros(height(t1),1); %colind will be the column index

for q = 1:height(t1)
    loc = find(ismember(vocabulary,lower(char(t1.t11(q)))));
    if isempty(loc)==1
        t1.colind(q)=0;
    else
        t1.colind(q) = loc ;

    end
end

rowt1 = t1.colind >0;
t1 = t1(rowt1,:);
if isempty(t1)== 1
 vec = zeros(1,15);
else
t1.Properties.VariableNames{1} = 'Term';
t1.Properties.VariableNames{2} = 'Frequency';
t1.Properties.VariableNames{3} = 'Percentage';
t1.Properties.VariableNames{4} = 'ColumnIndex';

for w =1:15
for e = 1:height(t1)
     t2(e,w) = t1.Frequency(e)*dtm((w),t1.ColumnIndex(e)) ;

end
end
      vec = sum (t2,1);

end

else
     vec = zeros(1,15);
end

else
     vec = zeros(1,15);
end
```

4.4.6 Boosting results using Levenshtein Distances

The document term matrix contains over 15 thousand text objects. These text objects include sequences of text and numbers. There is variation with the text object. Certain text objects are words that can be found in a dictionary, such as “pressure”, while certain text objects, like manufacturer name, serial number and model number cannot be found in an English dictionary but can be referenced from an industrial equipment lexicon which makes the metadata associated with each image (part of the image database). However, there is a third class of text objects which makes up the major proportion of the class. These include permutations of the first two classes. The MSER, OCR combination does a poor job when it comes to exact matches[71]. This can be attributed to the image quality attributes discussed in the subheading “Data” under the Problem Statement section. For instance, the word “pressure” has multiple permutations such as “pressur”, “presuresss” etc. The document term matrix was created with terms that occurred multiple times. Thus, we can be confident that each of these permutations does occur multiple times throughout the tag image database and is not an isolated occurrence. The encoding algorithm defined under “Encoding an image onto a text based feature space” encodes the image based only on the occurrence of exact matches. Thus, the weight of the permutations is completely ignored in the encoding process. For instance, let the term predicted by the NI-OCR function be “pressure”. ***Probability (Category$_i$\| pressure)*** is used to encode the image while ***Probability (Category$_i$\| permute[pressure])*** is ignored. While we (humans) can make the decision that “pressur” and “pressure” are the same and the missing leading “e” can be attributed to the output of the MSER [ ] text detection algorithm, a machine lacks the context to make that connection. To overcome this and generate context so that the algorithm might use the weights from the

permutations of the detected text we use a boosting algorithm that employs Levenshtein Distance [78].

Levenshtein distance is a string distance measure[78]. For single words, it can be defined as the number of operations (insertions, deletions and substitutions) that transform one string into another. For the example described, the distance between “pressure” and “pressur” would be 1. This gives the machine a measure to generate connections between the detected string and the terms in the document term matrix to identify reasonable permutations of the original string. However, the cut off distance (beyond what Levenshtein distance are the terms unrelated) needs to be either user defined or machine learned. We use the distribution of Levenshtein distances for a subset of terms to define the cutoff.

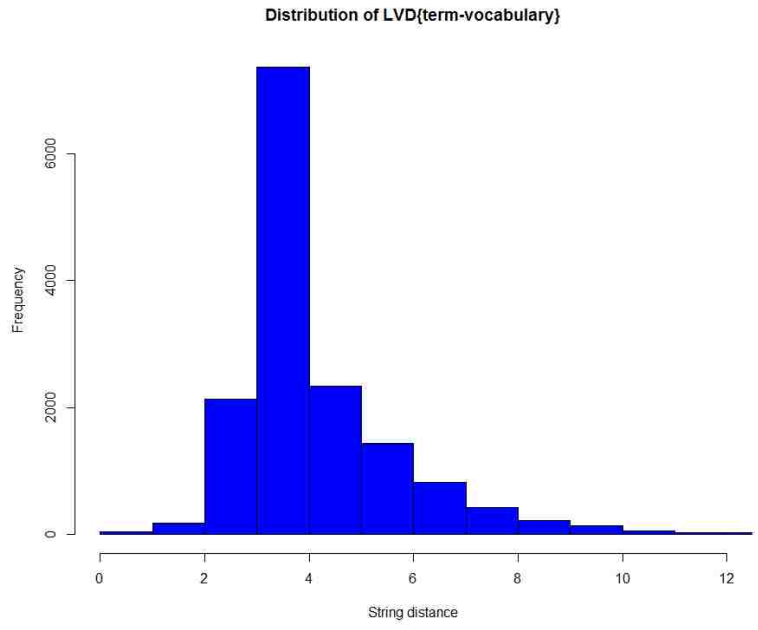


Figure 49 Distribution of String distances for sample term-[Vocabulary]

We need to determine the percentile for which the string distance is less than the cutoff. To determine this, a simulation was run for a subset with 3 percentiles, .05 %, 2.5% and 5%. The highest classification accuracy was achieved for .05% i.e. terms with string distances over .05 percentile are dropped. The weights of the remaining terms are used to encode the image.

Table 15: Document term matrix

	T1	T2	T3	T4	T5	T m
Cat1	.5	.3	.1	….	……	……..
Cat2	.2	.7	.2	…….	……..	………..
Cat3	.3	0	.7	……..	……..	……..
.	……..	……..	……..	……..	……..	……..
.	……..	……..	……..	……..	……..	……..
Cat 15	……..	……..	……..	……..	……..	……..

Table 16: Terms detected by NI-OCR

Term	Frequency of the m^{th} term
T_1	f_1
T_2	f_2
T_m	f_m

The document term matrix and the sample terms detected for an image are used as inputs for the mathematical formulation described below.

Mathematical Formulation

As defined previously, let the text based feature be defined by a vector V

$V = [V_1 + V_2 + V_3 + V_4 \ldots\ldots\ldots + V_n]$

Where, Vi defines the contribution of all terms detected to category i (classification category)

$V_1 = [f_1C(T_1) + f_2C(T_2) + f_3(T_3) \ldots\ldots\ldots + f_mC(T_m)]$

where, $f_1C(T_1)$ represents the contibution of T_1 to V_1

In the original encoding $C(T_1)$ was equal to A(1,1) which equals the probablity of T_1 appearing in an image representing category 1 .

We use a new way to define $C(T_1)$

$C(T_1) = f_1[\ w_1 A(1,t_1) + w_2 A(1,t_2) + w_k A(1,t_k) \ ]$

Where, $t = [t_1, t_2 \ldots\ldots\ldots\ldots t_k]$ are terms in the vocabulary such that

stringdist (T_1, t_k) < quantile (stringdist(T_1,[vocabulary]),0.0005) ie. terms with stringdist less than .05%

weights $w_i \propto 1/\text{stringdist}(T_1, t_i)$

4.4.7 Results

Table 17: Comparison to AlexNet

Machine Learning Algorithm	Hybrid Feature			Text feature			CNN Features		
	Accuracy	Prediction Speed(obs/sec)	Training Time(sec)	Accuracy	Prediction Speed(obs/sec)	Training Time(sec)	Accuracy	Prediction Speed(obs/sec)	Training Time(obs/sec)
Complex Tree	70.3	1400	104	69.1	280000	1.64	44.2	310	164.42
Medium Tree	40.5	2400	64	39.5	190000	2.3	34.6	2000	72
Simple Tree	25.7	1900	74	25.2	200000	0.62	21.1	1900	63.8
Linear Discriminant	76.6	240	2516	73.3	83000	2	67.5	470	1770
Quadratic Discriminant	NA	NA	NA	77.5	79000	1.84	NA	NA	NA
Linear SVM	84.9	35	2879.6	82	4900	21	83.1	60	3558
Quadratic SVM	87.1	5.5	4368	81.4	6900	61.7	85.4	14	4801
Cubic SVM	**87.4**	**5.5**	**5005**	**80.7**	**12000**	**146**	**85.8**	**4.8**	**5689**
SVM-Fine Guassian	17.3	5.7	6374	62	5400	27.75	17.1	4.2	8035
SVM-Medium Gaussian	86.1	5.8	5147	79.3	3600	17.4	84.9	16	4577
SVM-Coarse Gaussian	78.7	5.9	5197	70.7	3900	19.2	77.8	3.9	6401
Fine KNN	83.1	250	130	78.3	50000	0.805	82.5	120	161.44
Medium KNN	83.1	150	179.9	80.8	13000	4.635	82.4	250	205.85
Coarse KNN	74.5	220	235	81.4	6200	4.3266	73.6	75	322
Cosine KNN	84.9	220	268	81.1	5900	4.08	84.9	140	316
Cubic KNN	83.1	12	1157	80.4	2600	5.05	82.4	7.9	1614
Weighted KNN	83.9	120	392	81	38000	4.9	83.6	170	387.46
Ensemble Boosted Trees	65.6	170	2076	64.5	11000	17.26	59.5	380	1658.7
Ensemble Bagged Trees	**80.8**	**250**	**1572**	**83.7**	**14000**	**19.72**	**75.9**	**190**	**1881**
Ensemble Subspace Discriminant	**92.7**	**14**	**5901**	**72.6**	**2900**	**12.6**	**85.8**	**10**	**7550**
Ensemble Subspace KNN	83.2	19	2256	77.7	5100	10.3	82.9	16	2481
Ensemble RU Boosted Trees	40.5	620	3057	39.5	19000	17.35	34.6	520	3145
Average Measures	**71.90**	**383.26**	**2331.07**	**70.99**	**47109.09**	**18.29**	**68.08**	**317.23**	**2612.08**

Table 18: Comparison with VGG-VD19

Machine Learning Algorithm	Hybrid Feature			Text feature			CNN Features		
	Accuracy	Prediction Speed(obs/sec)	Training Time(sec)	Accuracy	Prediction Speed(obs/sec)	Training Time(sec)	Accuracy	Prediction Speed(obs/sec)	Training Time(obs/sec)
Complex Tree	69.5	5000	27.7	69.1	280000	1.64	54.3	9200	32.18
Medium Tree	39.3	11000	19.21	39.5	190000	2.3	47.8	11000	23.38
Simple Tree	25.2	13000	17.86	25.2	200000	0.62	24.8	12000	24.43
Linear Discriminant	**92.6**	610	53.8	73.3	83000	2	**85.6**	980	47.43
Quadratic Discriminant	NA	NA	NA	77.5	79000	1.84	NA	NA	NA
Linear SVM	85.1	260	711	82	4900	21	82.4	NA	NA
Quadratic SVM	87.3	111	890	81.4	6900	61.7	84.7	56	972
Cubic SVM	87.4	31	921.15	80.7	12000	146	85.3	48	1155
SVM-Fine Guassian	23	57	1270	62	5400	27.75	22.9	56	1369
SVM-Medium Gaussian	85.9	30	1142	79.3	3600	17.4	83.7	40	1251
SVM-Coarse Gaussian	79.2	29	1162	70.7	3900	19.2	77.6	60	1066
Fine KNN	81.5	1400	25.49	78.3	50000	0.805	80.5	980	39.4
Medium KNN	82.7	770	39.54	80.8	13000	4.635	81.7	1000	47.48
Coarse KNN	75.3	1000	45.2	81.4	6200	4.3266	74.5	580	69.5
Cosine KNN	82.2	300	74.35	81.1	5900	4.08	81.7	820	62.345
Cubic KNN	83.3	41	298.84	80.4	2600	5.05	82	35	343.7
Weighted KNN	83.1	480	74	81	38000	4.9	82.2	910	79.4
Ensemble Boosted Trees	68.2	1200	478.61	64.5	11000	17.26	63.3	1200	814
Ensemble Bagged Trees	84	2800	324.23	**83.8**	14000	19.72	77.3	2800	145
Ensemble Subspace Discriminant	91.5	82	819.34	72.6	2900	12.6	85.1	74	787
Ensemble Subspace KNN	82.5	97	460	77.7	5100	10.3	80.8	91	445
Ensemble RU Boosted Trees	39.3	2200	635.05	39.5	19000	17.35	47.8	2300	830
Average Measures	**72.77**	**1928.48**	**451.87**	**70.99**	**47109.09**	**18.29**	**70.76**	**2211.50**	**480.16**

Table 19: Comparison to SURF

Machine Learning Algorithm	Hybrid Feature			Text feature			CNN Features		
	Accuracy	Prediction Speed(obs/sec)	Training Time(sec)	Accuracy	Prediction Speed(obs/sec)	Training Time(sec)	Accuracy	Prediction Speed(obs/sec)	Training Time(obs/sec)
Complex Tree	70	11000	15.37	69.1	280000	1.64	25.4	11000	16.4
Medium Tree	41.4	11000	10.773	39.5	190000	2.3	20.9	14000	11.3
Simple Tree	25.9	12000	9.91	25.2	200000	0.62	15.6	6800	9.471
Linear Discriminant	82.9	200	85.191	73.3	83000	2	56.1	320	64.135
Quadratic Discriminant	NA	NA	NA	77.5	79000	1.84	FAILED	NA	NA
Linear SVM	70.4	300	696.3	82	4900	21	59.2	320	712
Quadratic SVM	71.4	20	1184	81.4	6900	61.7	**60.8**	18	1203.7
Cubic SVM	70.7	33	883.31	80.7	12000	146	60.4	36	892
SVM-Fine Guassian	12.8	27	1307	62	5400	27.75	13	27	1318
SVM-Medium Gaussian	66.2	26	1295.2	79.3	3600	17.4	57.7	25	1306
SVM-Coarse Gaussian	58.2	35	905.12	70.7	3900	19.2	49.4	37	908
Fine KNN	43	670	22.9	78.3	50000	0.805	35.3	810	22.93
Medium KNN	34.3	970	30.423	80.8	13000	4.635	30.5	1000	30.615
Coarse KNN	27.1	1300	36.4	81.4	6200	4.3266	21.4	1200	34.234
Cosine KNN	58.1	770	50.854	81.1	5900	4.08	52.5	750	51.157
Cubic KNN	40.2	68	191.95	80.4	2600	5.05	35.8	76	180.68
Weighted KNN	36.6	670	73.208	81	38000	4.9	32.2	660	75.85
Ensemble Boosted Trees	62.9	2400	358.65	64.5	11000	17.26	23.6	2200	265.4
Ensemble Bagged Trees	71.2	1300	205.57	**83.8**	14000	19.72	41.7	1600	263.37
Ensemble Subspace Discriminant	**83.8**	85	574.8	72.6	2900	12.6	59	81	551.38
Ensemble Subspace KNN	83.8	79	35561	77.7	5100	10.3	43.5	83	414.13
Ensemble RU Boosted Trees	41.4	2000	454.88	39.5	19000	17.35	20.9	1900	473.89
Average Measures	54.87	2140.62	2092.99	70.99	47109.09	18.29	38.80	2044.90	419.27

Table 20: Cluster Size vs. Test Accuracy: SIFT based bag of features

	Testing Accuracy based on a 50 % holdout validation					
Test ID	Images per category	Cluster size	Strongest Features	Training Accuracy	Testing Accuracy	Time (min)
1	500	10000	0.99	0.96	0.6	44
2	500	20000	0.99	0.98	0.6	70
3	750	10000	0.99	0.95	0.58	73
4	750	20000	0.99	0.98	0.63	111
5	750	40000	0.99	Fail	Fail	Fail
6	750	30000	0.99	0.98	0.63	144
7	900	10000	0.99	0.95	0.62	94
8	970	30000	0.99	0.98	0.66	190

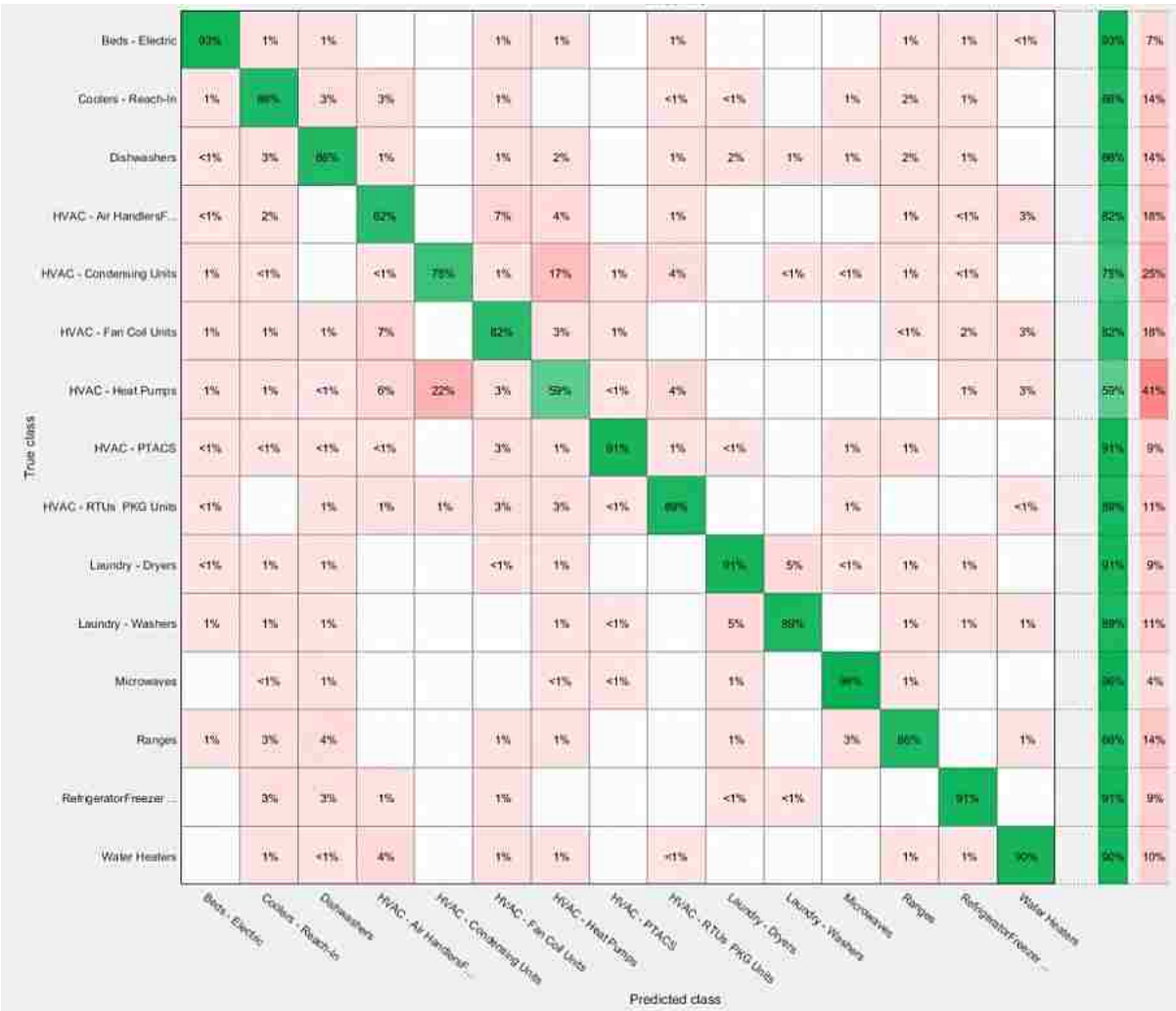


Figure 50: Cubic SVM trained on Image features-AlexNet | True positive-False Negative|Accuracy 85.8 %

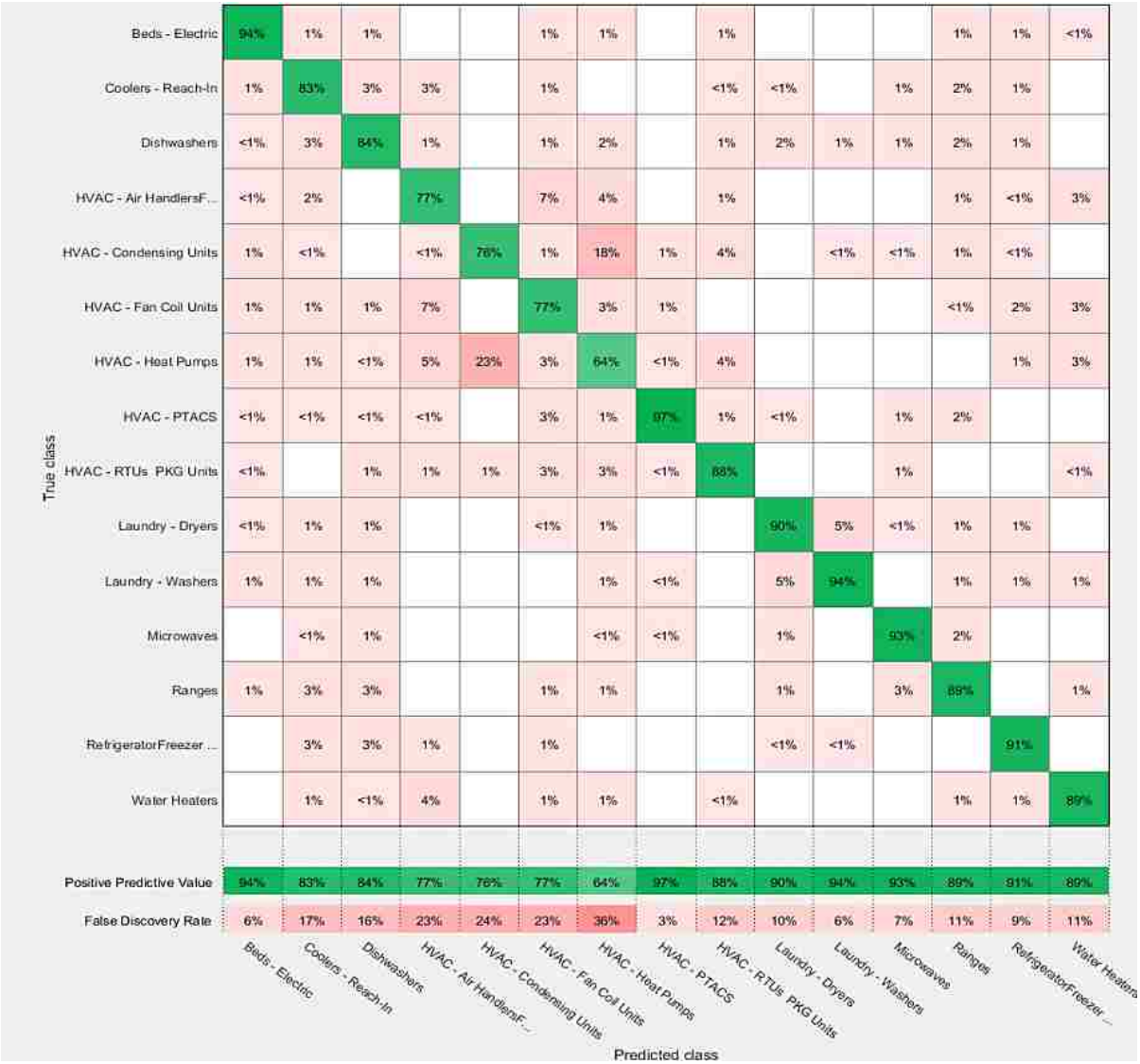


Figure 51: Cubic SVM trained on Image features-AlexNet | Positive Prediction-False Discovery Rate|Accuracy 85.8 %

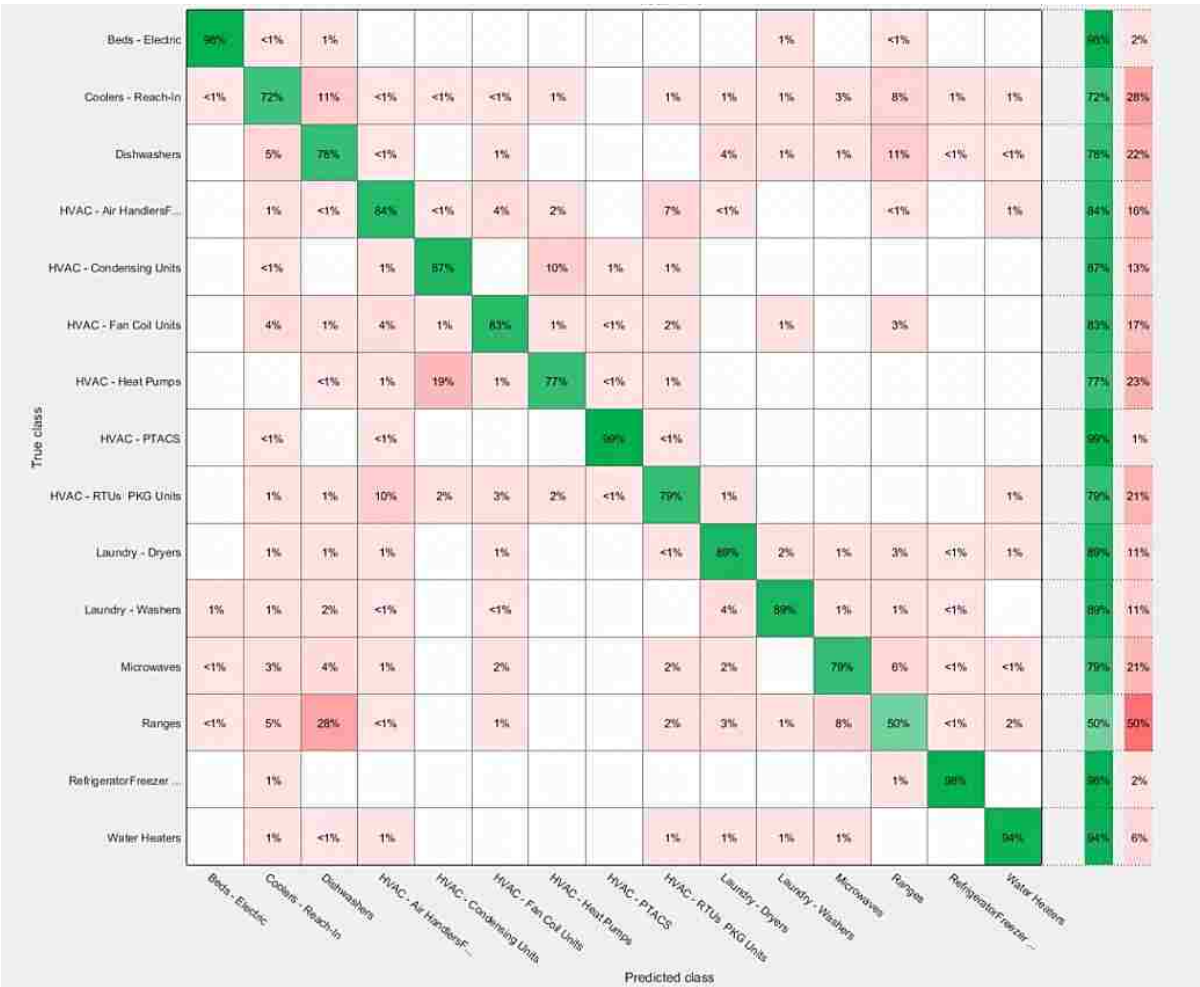


Figure 52: Ensemble Bagged Tree trained on Text features | True Positive-False Negative|Accuracy 83.7 %

Figure 53: Ensemble Bagged Tree trained on Text features | Positive Prediction-False Discovery Rate |Accuracy 83.7 %

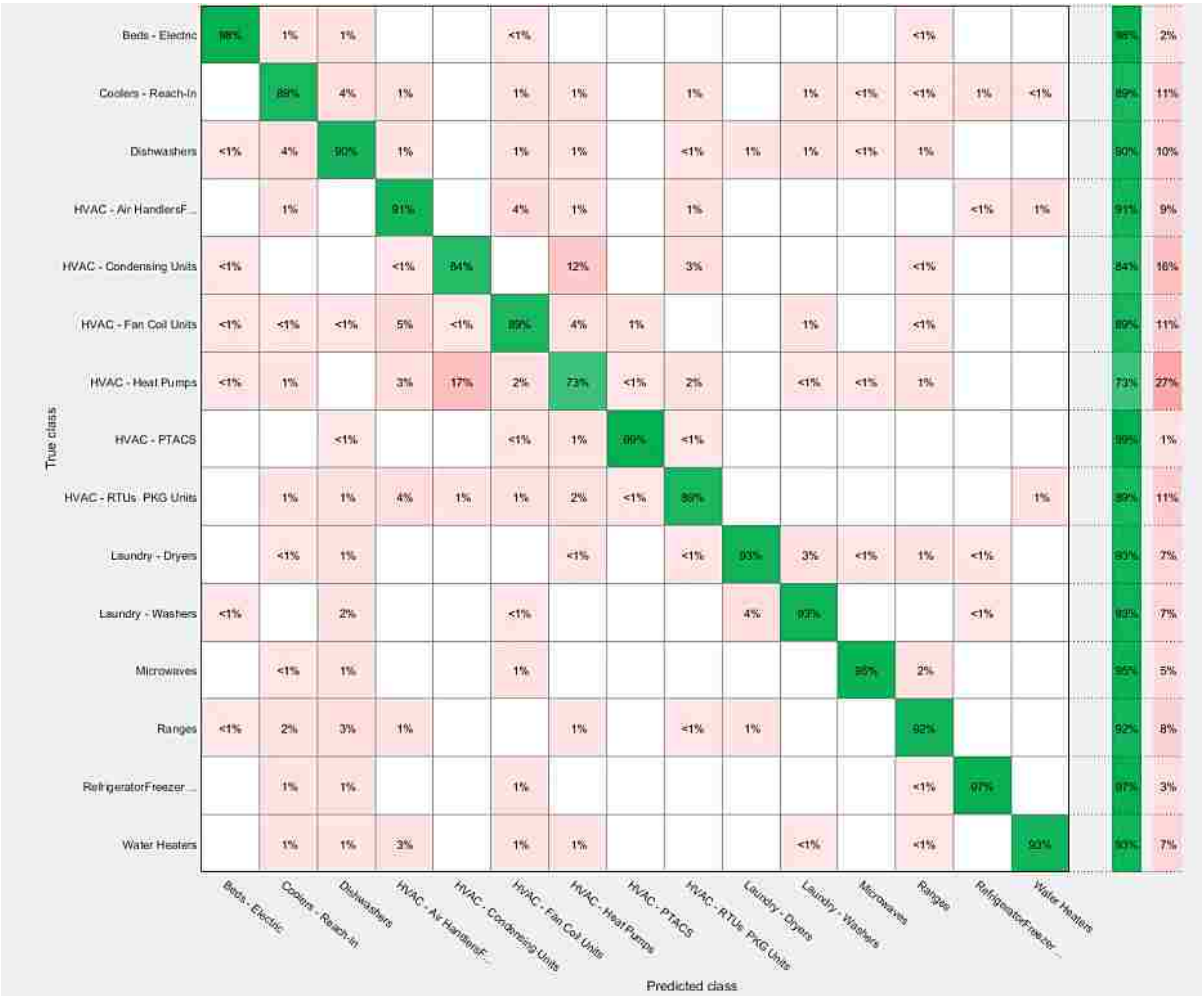


Figure 54: Ensemble Subspace Discriminant trained on Combined Features (Alexnet) | True Positive-False Negative|Accuracy 92.7 %

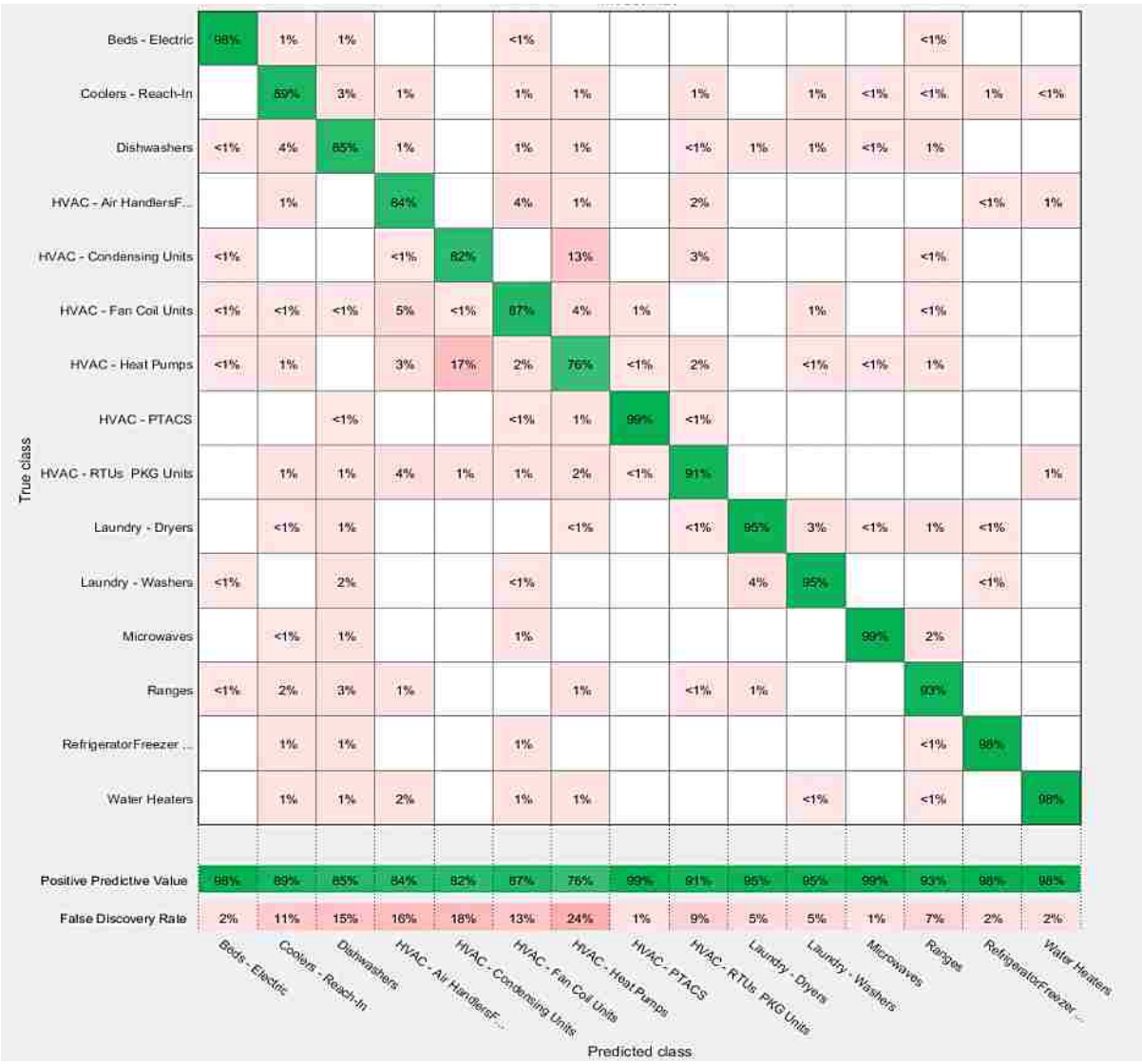


Figure 55: Ensemble Subspace Discriminant trained on Combined Features (Alexnet) | Positive Prediction-False Discovery Rate |Accuracy 92.7 %

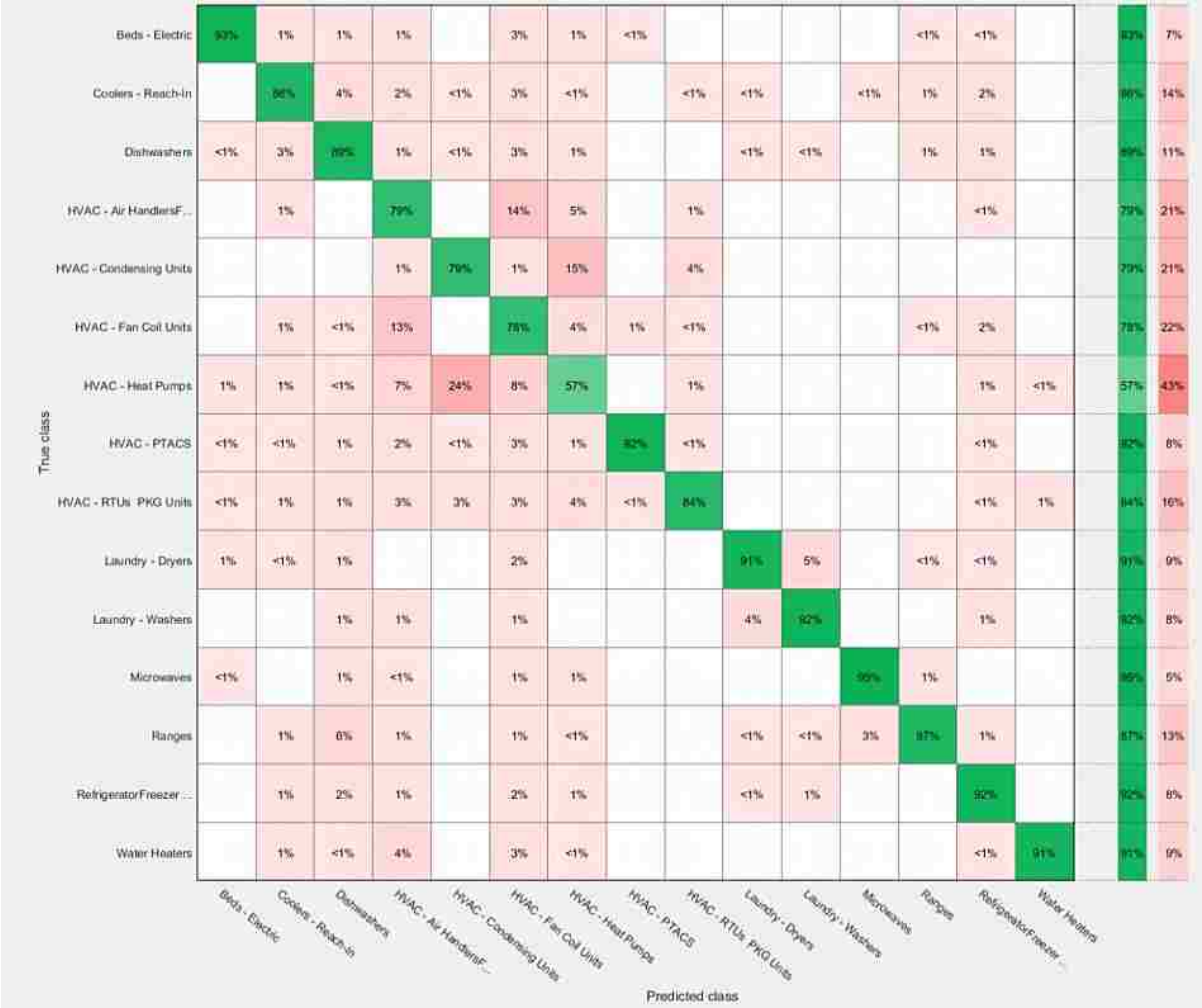


Figure 56: Linear Discriminant trained on Image features-VGG-VD19 | True Positive-False Negative|Accuracy 85.6 %

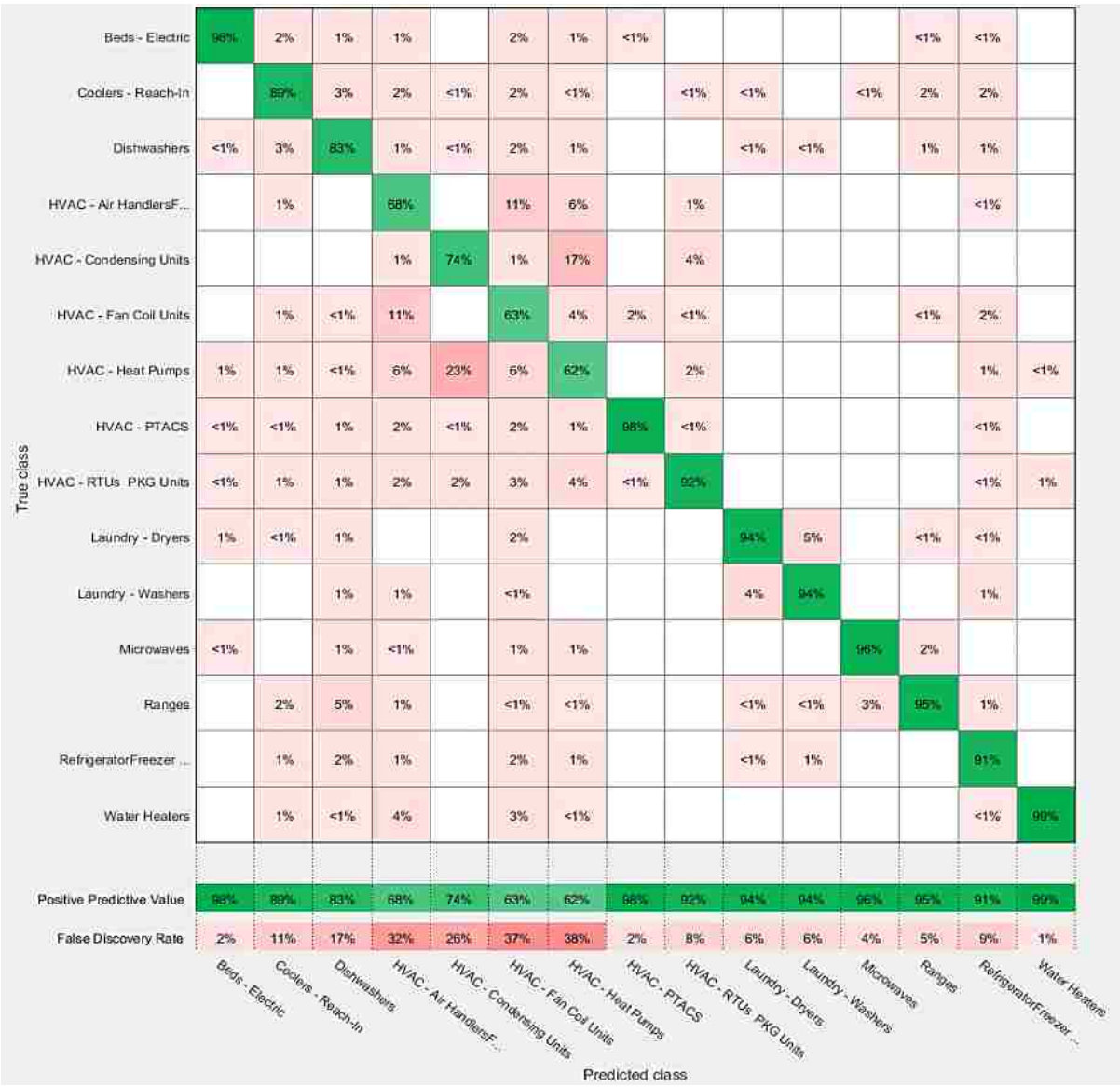


Figure 57: Linear Discriminant trained on Image features-VGG-VD19 | Positive Prediction-False Discovery Rate |Accuracy 85.6 %

Figure 58: Linear Discriminant trained on Combined Features (VGG-VD) | True positive-False Negative|Accuracy 92.6 %

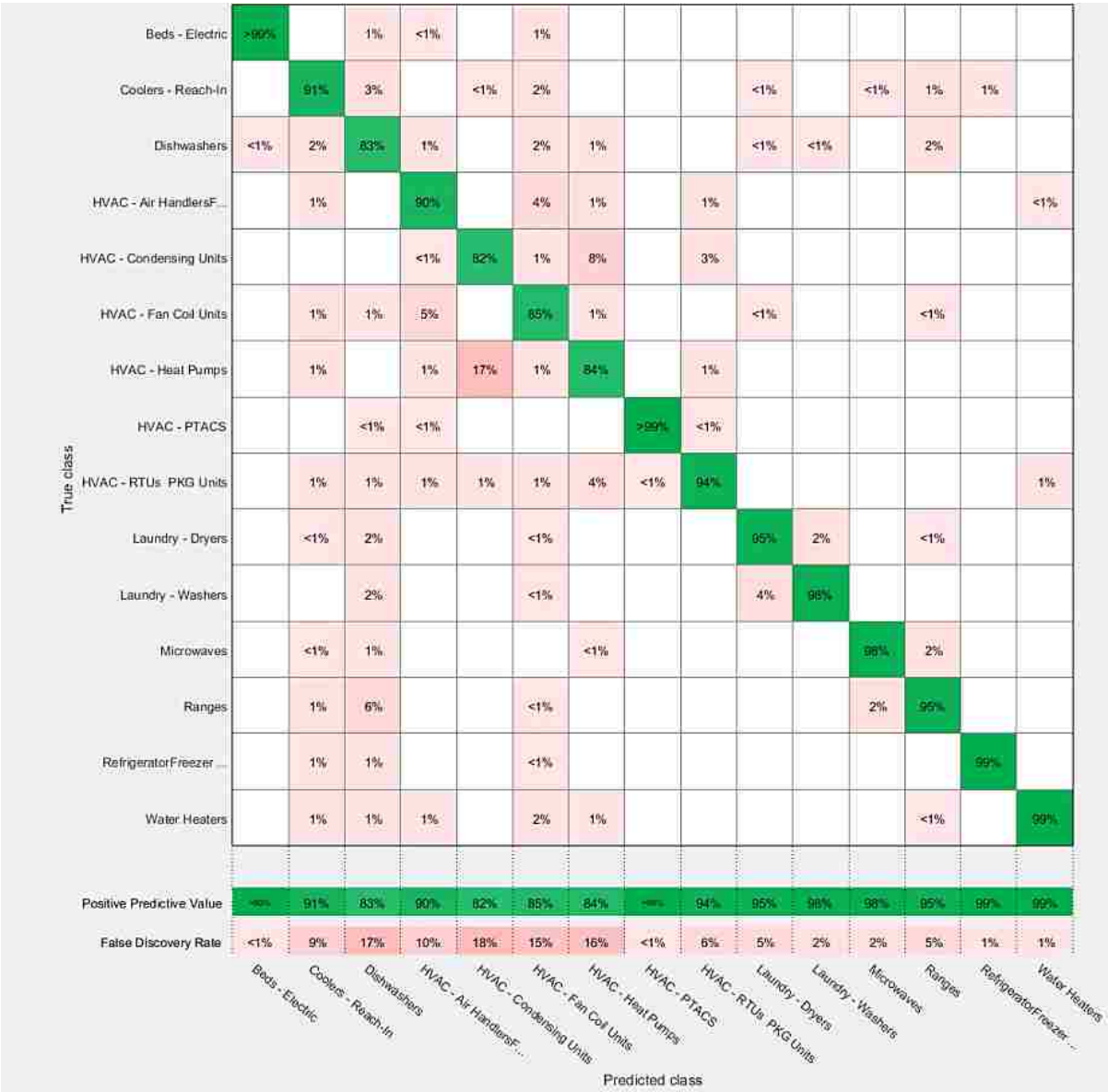


Figure 59: Linear Discriminant trained on Combined Features (VGG-VD) | Positive prediction-False Discovery Rate |Accuracy 92.6 %

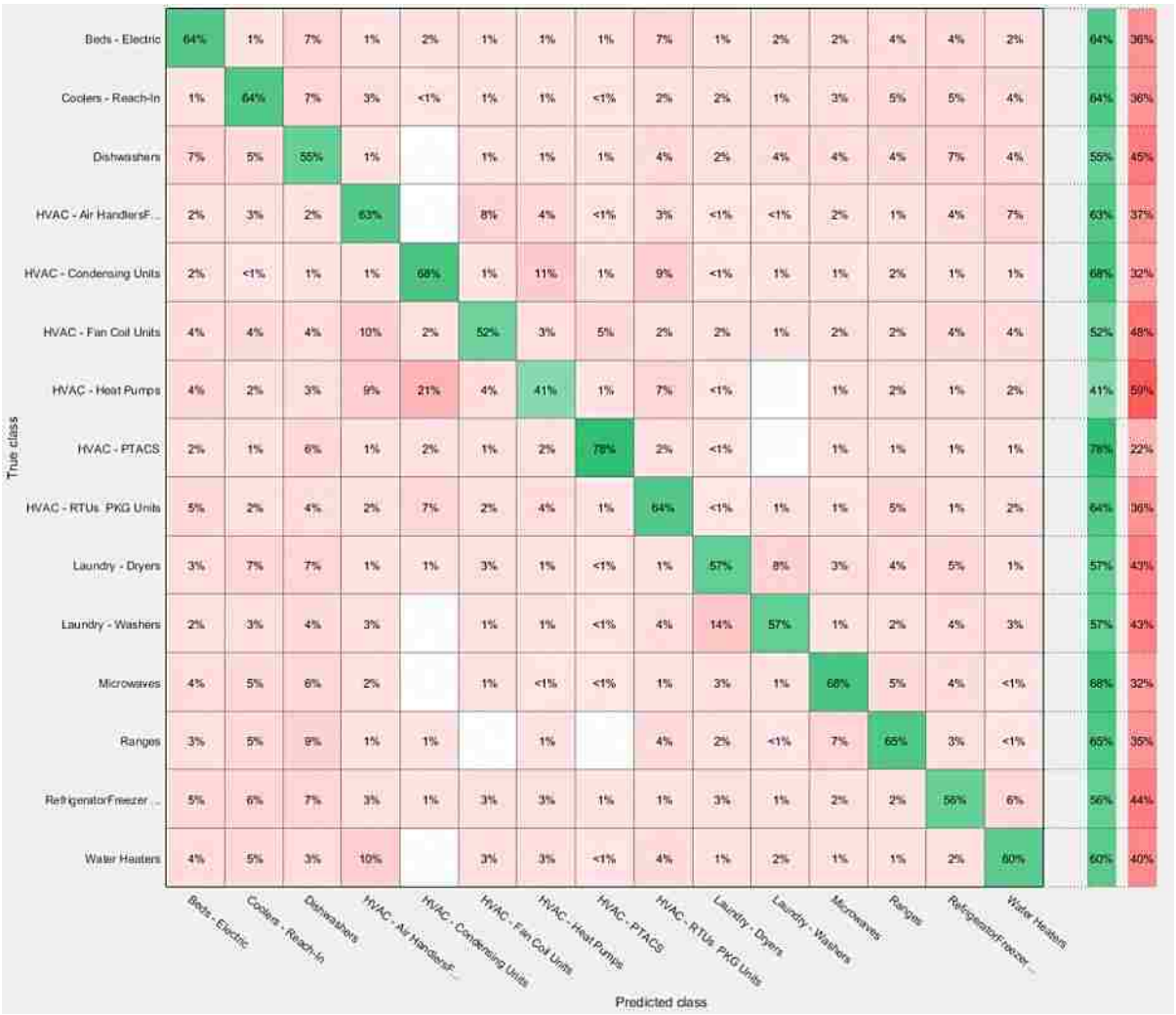


Figure 60: Quadratic SVM trained on Image features-SURF | True Positive-False Negative|Accuracy 60.8 %

Figure 61: Quadratic SVM trained on Image features-SURF | Positive prediction-False Discovery Rate |Accuracy 60.8 %

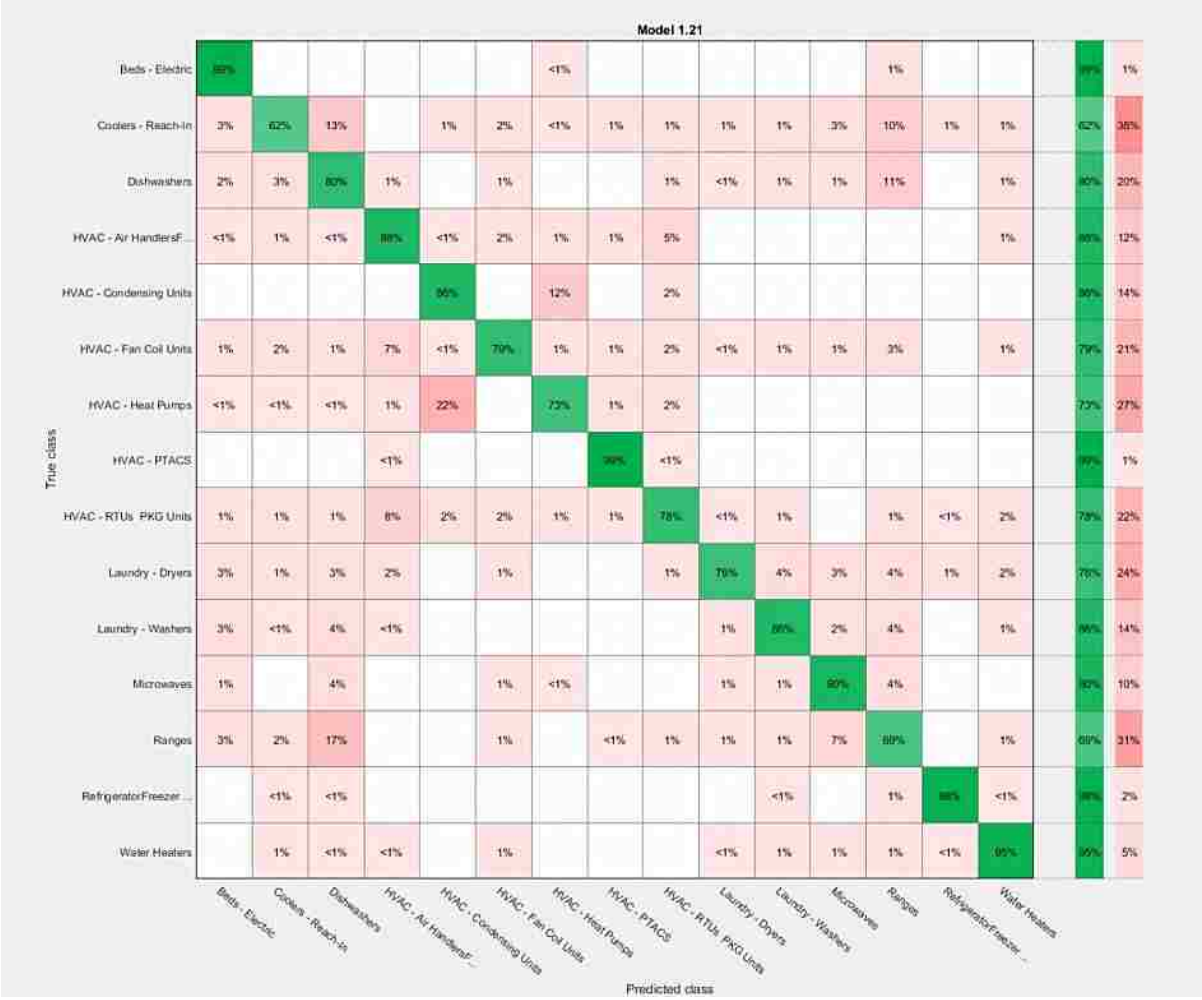


Figure 62: Ensemble Subspace KNN trained on Combined Features (SURF) | True positive-False Negative|Accuracy 83.8 %

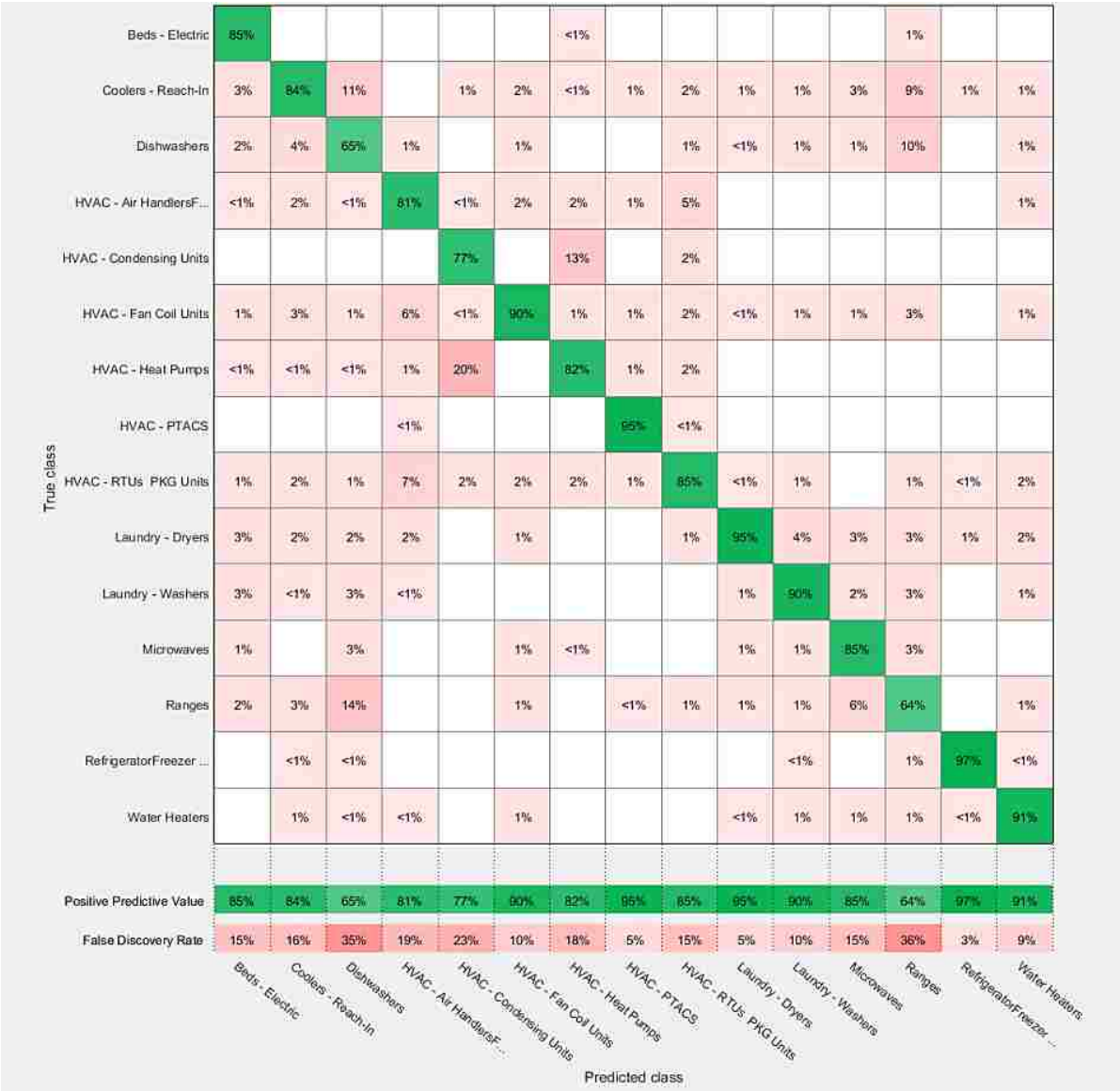


Figure 63: Ensemble Subspace KNN trained on Combined Features (SURF) | Positive Prediction-False Discovery Rate |Accuracy 83.8 %

4.5 Evaluation

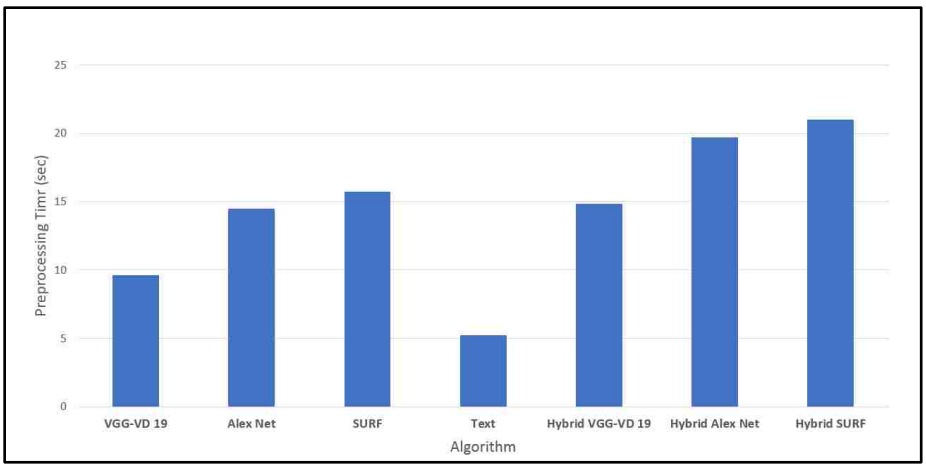
The algorithms are compared by comparing the prediction accuracy of the of the classifiers trained in the respective feature space's. These are provided in 4.4.7, Results.

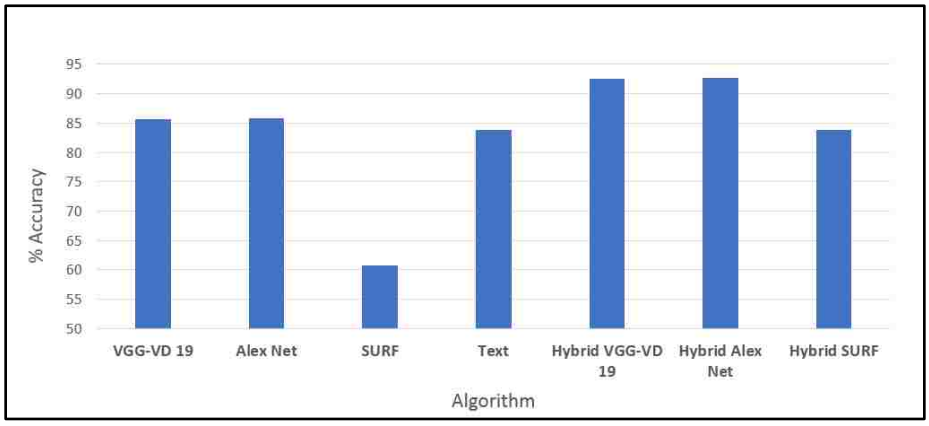
4.5.1 Comparing our feature space to state of the art

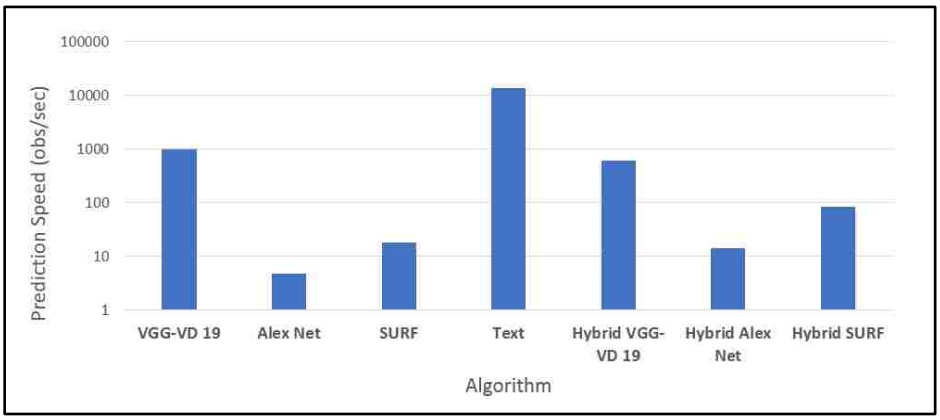
The dataset used to create this table includes over 20000 images evenly spread across 15 classification categories. The accuracy was obtained using a 50 percent holdout validation (25 percent holdout validation, 5-fold cross validation, 10-fold cross validation and 20-fold cross validation were also conducted.) since it was the most conservative estimate. It must also be noted that the accuracy presented in the table below is the highest accuracy obtained using 22 separate machine learning classifiers. For more information please refer to *Table 17,Table 18,Table 19* . For more background on the machine learning classifiers please refer to the appendix.

Table 21: Comparison table

	Accuracy	Prediction Speed (obs/sec)	Training Time (sec)	Preprocessing time (sec/image)
VGG-VD 19	85.6	980	47.43	9.61
Alex Net	85.8	4.8	5689	14.45
SURF	60.8	18	1203.7	15.73
Text	83.8	14000	19.72	5.24
Hybrid VGG-VD 19	92.6	610	53.8	14.85
Hybrid Alex Net	92.7	14	5901	19.69
Hybrid SURF	83.8	85	574	20.97


25
20
15
10
5
0
Preprocessing Timr (sec)
VGG-VD 19
Alex Net
SURF
Text
Hybrid VGG-VD 19
Hybrid Alex Net
Hybrid SURF
Algorithm


95
90
85
80
75
70
65
60
55
50
% Accuracy
VGG-VD 19
Alex Net
SURF
Text
Hybrid VGG-VD 19
Hybrid Alex Net
Hybrid SURF
Algorithm


100000
10000
1000
100
10
1
Prediction Speed (obs/sec)
VGG-VD 19
Alex Net
SURF
Text
Hybrid VGG-VD 19
Hybrid Alex Net
Hybrid SURF
Algorithm

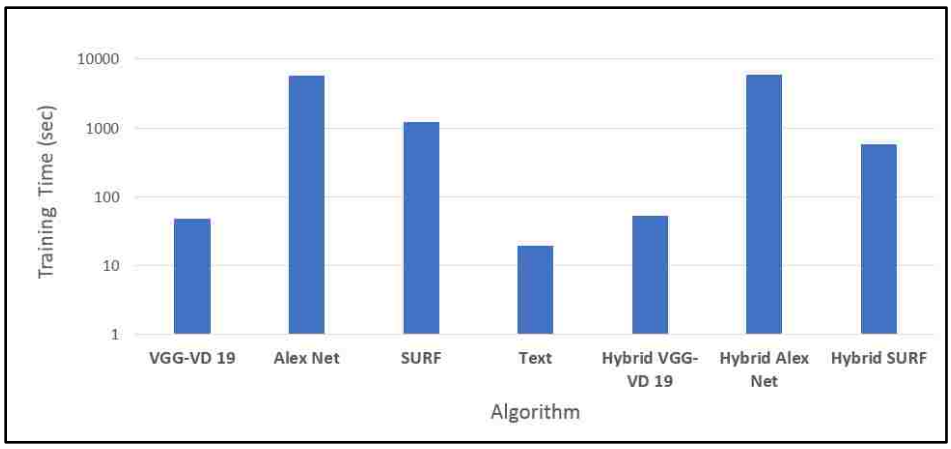


Figure 64: Comparing our work to the state of the art

5.0 CONCLUSION

5.1 Contributions

1. In this research, we present and test the feature space augmentation techniques to imitate the human behavior associated with inductive and deductive reasoning. The human enables iterative learning framework (a subset of longitudinal feature space augmentation) is a novel concept.
2. In the emotion recognition (longitudinal FSA) problem we present 3 models,
    i. 6 category dominant state model: The accuracy we achieve is the highest amongst related works (87.4%).
    ii. 3 category classification model: We achieve the highest accuracy amongst the related works. Moreover, we establish for the first time that there is no information gain from introducing PPG in a 3 category classification (92%).
    iii. Emotional Spectrum is a novel concept introduced and validated in this work.

Moreover, the dataset curated is the largest “stimuli-emotion induction” database in the scientific field.

3. The image recognition problem (latitudinal FSA) introduces a novel robust feature that matches the accuracy of the state of the art CNN’s and is significantly less compute intensive. The compact feature provides statistically significant information gain when added to CNN based image features.

5.2 Broader Impact

This work is aligned with the Marquette sprit of “Magis” and Service. It is our privilege at this fine institution to advance the body of science that has an impact on the world. Our work in Autism and Image recognition were both inspired by the goal of improving lives of members in our community. It is our responsibility to embrace “Cura Personalis”. I have faith that the work presented in this dissertation will have a meaningful impact on the community.

5.2.1 Short Term

The system (emotion recognition) designed will be implemented at the PEERS intervention at Marquette University. In the past we have struggled to scale the computational aspect to the entire class due to the cost of the sensors. However, PPG sensors are 1/25 the cost of the E4 sensors currently being used. This would allow is to equip the entire class with wearable sensors. The human enabled iterative learning framework (longitudinal FSA) will allow us to tailor the model to specific individuals. An emotional state dashboard will significantly reduce the trained personnel required to run such interventions and allow for larger class sizes. It will make intervention cheaper and accessible to more individuals. The image classifier (problem 2) will be implemented in a network independent mobile application.

5.2.2 Long Term

Over a longer period, we expect the (emotion detection) system (human enabled iterative learning, the data collection system, dominant state model and the emotional spectrum

model) will be expanded to other mental disorders such as post-traumatic stress disorder (PTSD), Schizophrenia and Hypertension.

5.3 Future Work

This work will be expanded by the new graduate students at the ubicomp lab to implement real time systems for emotion modelling and recognition in ASD populations and veterans, both of which are current lab collaborations. The image classification algorithm will be implemented as a part of a larger asset management program at Direct Supply.

6.0 REFERENCES

[1] J. Bosch and H. H. Olsson, "Data-driven continuous evolution of smart systems," in *Proceedings of the 11th International Symposium on Software Engineering for Adaptive and Self-Managing Systems,* 2016, pp. 28-34.

[2] M. Endler *et al*, "Towards Stream-based Reasoning and Machine Learning for IoT Applications," 2017.

[3] M. J. Gajjar, *Mobile Sensors and Context-Aware Computing.* Morgan Kaufmann, 2017.

[4] S. Kounev *et al*, "The notion of self-aware computing," in *Self-Aware Computing Systems*Anonymous Springer, 2017, pp. 3-16.

[5] H. H. Olsson and J. Bosch, "From opinions to data-driven software r&d: A multi-case study on how to close the'open loop'problem," in *Software Engineering and Advanced Applications (SEAA), 2014 40th EUROMICRO Conference On,* 2014, pp. 9-16.

[6] (). *MILWAUKEE PEERS PROJECT*. Available: http://www.marquette.edu/psyc/about_PEERS.shtml.

[7] (). *Ubicomp Lab*. Available: http://ubicomp.mscs.mu.edu.

[8] Zablotsky B, Black LI, Maenner MJ, et al., "Estimated Prevalence of Autism and Other

Developmental Disabilities Following Questionnaire

Changes in the 2014 National Health Interview Survey," *National Health Statistics Reports (NHSR),* vol. 87, 2015.

[9] (). *Facts and Statistics *. Available: http://www.autism-society.org/what-is/facts-and-statistics/.

[10] J. Leigh and J. Du, "Brief Report: Forecasting the Economic Burden of Autism in 2015 and 2025 in the United States," *J Autism Dev Disord,* vol. 45, *(12),* pp. 4135-4139, 2015. Available: http://www.ncbi.nlm.nih.gov/pubmed/26183723. DOI: 10.1007/s10803-015-2521-7.

[11] (). *High-Quality Early Intervention for Autism More Than Pays for Itself*. Available: https://www.autismspeaks.org/science/science-news/high-quality-early-intervention-autism-more-pays-itself.

[12] Y. Nah, R. Young and N. Brewer, "Using the Autism Detection in Early Childhood (ADEC) and Childhood Autism Rating Scales (CARS) to Predict Long Term Outcomes in Children with Autism Spectrum Disorders," *J Autism Dev Disord,* vol. 44, *(9),* pp. 2301-

2310, 2014. Available: http://www.ncbi.nlm.nih.gov/pubmed/24658894. DOI: 10.1007/s10803-014-2102-1.

[13] A. V. S. Buescher *et al*, "Costs of Autism Spectrum Disorders in the United Kingdom and the United States," *JAMA Pediatrics,* vol. 168, *(8),* pp. 721-728, 2014. Available: http://dx.doi.org/10.1001/jamapediatrics.2014.210. DOI: 10.1001/jamapediatrics.2014.210.

[14] K. Järbrink, "The economic consequences of autistic spectrum disorder among children in a Swedish municipality," *Autism,* vol. 11, *(5),* pp. 453-463, 2007. Available: http://journals.sagepub.com/doi/full/10.1177/1362361307079602. DOI: 10.1177/1362361307079602.

[15] P. Barros *et al*, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," *Neural Networks,* vol. 72, pp. 140-151, 2015.

[16] G. Valenza and E. P. Scilingo, "Conclusions and discussion on mood and emotional-state recognition using the autonomic nervous system dynamics," in *Autonomic Nervous System Dynamics for Mood and Emotional-State Recognition*Anonymous Springer, 2014, pp. 127-138.

[17] L. Likforman-Sulem *et al*, "EMOTHAW: A Novel Database for Emotional State Recognition From Handwriting and Drawing," *IEEE Transactions on Human-Machine Systems,* vol. 47, *(2),* pp. 273-284, 2017.

[18] M. Neji *et al*, "Towards an intelligent information research system based on the human behavior: Recognition of user emotional state," in 2013, pp. 371-376.

[19] X. Wang, D. Nie and B. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing,* vol. 129, pp. 94-106, 2014.

[20] D. Filko and G. Martinović, "Emotion recognition system by a neural network based facial expression analysis," *Automatika,* vol. 54, *(2),* pp. 263-272, 2013.

[21] A. Agrawal and N. K. Mishra, "Fusion based emotion recognition system," in 2016, pp. 727-732.

[22] R. Jiang *et al*, "Emotion recognition from scrambled facial images via many graph embedding," *Pattern Recognit,* vol. 67, pp. 245-251, 2017.

[23] (). *E4 Wristband*. Available: https://www.empatica.com/e4-wristband.

[24] M. Izumi *et al*, "Changes in autonomic nervous system activity, body weight, and percentage fat mass in the first year postpartum and factors regulating the return to pre-pregnancy weight," *Journal of Physiological Anthropology,* vol. 35, *(1),* pp. 26, 2016.

[25] E. Roggero *et al*, "The sympathetic nervous system affects the susceptibility and course of Trypanosoma cruzi infection," *Brain Behav. Immun.,* vol. 58, pp. 228-236, 2016.

[26] K. Ludwig *et al*, "The autonomic nervous system," in *NEUROPROSTHETICS: Theory and Practice*Anonymous World Scientific, 2017, pp. 12-39.

[27] M. A. Morrison *et al*, "Studying the peripheral sympathetic nervous system and neuroblastoma in zebrafish," *Methods Cell Biol.,* vol. 134, pp. 97-138, 2016.

[28] M. Ragot *et al*, "Emotion recognition using physiological signals: Laboratory vs. wearable sensors," in 2017, pp. 15-22.

[29] (). *BIOPAC MP-150*. Available: https://www.biopac.com/product/mp150-data-acquisition-systems/.

[30] P. J. Lang, M. M. Bradley and B. N. Cuthbert, "International affective picture system (IAPS): Technical manual and affective ratings," *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida,* vol. 2, 1999.

[31] K. Nisa'Minhad *et al*, "Human emotion classifications for automotive driver using skin conductance response signal," in 2016, pp. 371-375.

[32] C. He, Y. Yao and X. Ye, "An emotion recognition system based on physiological signals obtained by wearable sensors," in *Wearable Sensors and Robots*Anonymous Springer, 2017, pp. 15-25.

[33] C. Maaoui and A. Pruski, "Emotion recognition through physiological signals for human-machine communication," in *Cutting Edge Robotics 2010*Anonymous InTech, 2010, .

[34] R. W. Picard, E. Vyzas and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 23, *(10),* pp. 1175-1191, 2001.

[35] A. Haag *et al*, "Emotion recognition using bio-sensors: First steps towards an automatic system," in 2004, pp. 36-48.

[36] J. P. Arias, C. Busso and N. B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Comput. Speech Lang.,* vol. 28, *(1),* pp. 278-294, 2014.

[37] Z. Zhang *et al*, "Emotion detection using kinect 3D facial points," in 2016, pp. 407-410.

[38] M. Soleymani *et al*, "Continuous emotion detection using EEG signals and facial expressions," in 2014, pp. 1-6.

[39] C. Loconsole *et al*, "Real-time emotion recognition novel method for geometrical facial features extraction," in 2014, pp. 378-385.

[40] M. Liu *et al*, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in 2014, pp. 494-501.

[41] A. Veenendaal *et al*, "Group Emotion Detection using Edge Detecttion Mesh Analysis," *Computer Science and Emerging Research Journal,* vol. 2, 2014.

[42] R. Rakshit, V. R. Reddy and P. Deshpande, "Emotion detection and recognition using HRV features derived from photoplethysmogram signals," in 2016, pp. 2.

[43] Y. Rao *et al*, "Affective topic model for social emotion detection," *Neural Networks,* vol. 58, pp. 29-37, 2014.

[44] J. Lei *et al*, "Towards building a social emotion detection system for online news," *Future Generation Comput. Syst.,* vol. 37, pp. 438-448, 2014.

[45] F. Yu *et al*, "Emotion detection from speech to enrich multimedia content," *Advances in Multimedia Information processing—PCM 2001,* pp. 550-557, 2001.

[46] J. Hewig *et al*, "Brief report," *Cognition & Emotion,* vol. 19, *(7),* pp. 1095-1109, 2005.

[47] M. Garbarino *et al*, "Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition," in *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference On,* 2014, pp. 39-42.

[48] R. J. Povinelli *et al*, "Time series classification using Gaussian mixture models of reconstructed phase spaces," *IEEE Trans. Knowled. Data Eng.,* vol. 16, *(6),* pp. 779-783, 2004.

[49] J. Ye, R. J. Povinelli and M. T. Johnson, "Phoneme classification using naive bayes classifier in reconstructed phase space," in *Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop. Proceedings of 2002 IEEE 10th,* 2002, pp. 37-40.

[50] G. Lera and M. Pinzolas, "Neighborhood based Levenberg-Marquardt algorithm for neural network training," *IEEE Trans. Neural Networks,* vol. 13, *(5),* pp. 1200-1203, 2002.

[51] D. E. Budil *et al*, "Nonlinear-least-squares analysis of slow-motion EPR spectra in one and two dimensions using a modified Levenberg–Marquardt algorithm," *Journal of Magnetic Resonance, Series A,* vol. 120, *(2),* pp. 155-189, 1996.

[52] M. I. Lourakis, "A brief description of the Levenberg-Marquardt algorithm implemented by levmar," *Foundation of Research and Technology,* vol. 4, *(1),* 2005.

[53] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences,* vol. 79, *(8),* pp. 2554-2558, 1982.

[54] D. E. Rumelhart, G. E. Hinton and J. L. McClelland, "A general framework for parallel distributed processing," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* vol. 1, pp. 45-76, 1986.

[55] Y. LeCun *et al*, "Gradient-based learning applied to document recognition," *Proc IEEE,* vol. 86, *(11),* pp. 2278-2324, 1998.

[56] L. Deng, "The MNIST database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Process. Mag.,* vol. 29, *(6),* pp. 141-142, 2012.

[57] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol. (Lond.),* vol. 160, *(1),* pp. 106-154, 1962.

[58] (). *A Beginner's Guide To Understanding Convolutional Neural Networks*. Available: https://adeshpande3.github.io/adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/.

[59] (). *ImageNet*. Available: http://www.image-net.org/.

[60] A. Vedaldi and K. Lenc, "MatConvNet," in Oct 13, 2015, pp. 689-692.

[61] J. Donahue *et al*, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International Conference on Machine Learning,* 2014, pp. 647-655.

[62] A. Sharif Razavian *et al*, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,* 2014, pp. 806-813.

[63] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems,* 2012, pp. 1097-1105.

[64] (Jul 15,). *AlexNet Visualization*. Available: https://jeremykarnowski.wordpress.com/2015/07/15/alexnet-visualization/.

[65] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv Preprint arXiv:1409.1556,* 2014.

[66] G. Csurka *et al*, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV,* 2004, pp. 1-2.

[67] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. the Proceedings of the Seventh IEEE International Conference On,* 1999, pp. 1150-1157.

[68] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, *(2),* pp. 91-110, 2004.

[69] H. Bay, T. Tuytelaars and L. Van Gool, "Surf: Speeded up robust features," *Computer vision–ECCV 2006,* pp. 404-417, 2006.

[70] (04/10/17). *Image Classification with Bag of Visual Words*. Available: https://www.mathworks.com/help/vision/ug/image-classification-with-bag-of-visual-words.html.

[71] J. Matas *et al*, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vision Comput.,* vol. 22, *(10),* pp. 761-767, 2004.

[72] (04/10/17). *Region Detectors *. Available: http://www.micc.unifi.it/delbimbo/wp-content/uploads/2011/03/slide_corso/A34%20MSER.pdf.

[73] H. Chen *et al*, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *Image Processing (ICIP), 2011 18th IEEE International Conference On,* 2011, pp. 2609-2612.

[74] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference On,* 2012, pp. 3538-3545.

[75] A. Gonzalez *et al*, "Text location in complex images," in *Pattern Recognition (ICPR), 2012 21st International Conference On,* 2012, pp. 617-620.

[76] Y. Li and H. Lu, "Scene text detection via stroke width," in *Pattern Recognition (ICPR), 2012 21st International Conference On,* 2012, pp. 681-684.

[77] (04/10/17). *regionprops*. Available: https://www.mathworks.com/help/images/ref/regionprops.html.

[78] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet Physics Doklady,* 1966, pp. 707-710.

7.0 BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv Preprint arXiv:1603.04467,*

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Isard, M. (2016). TensorFlow: A system for large-scale machine learning. *, 16* 265-283.

Ackermann, P., Kohlschein, C., Bitsch, J. Á, Wehrle, K., & Jeschke, S. (2016). EEG-based automatic emotion recognition: Feature extraction, selection and classification methods. 1-6.

Adorni, G., Cagnoni, S., Gori, M., & Maratea, M. (2016). *AI* IA 2016 advances in artificial intelligence: XVth international conference of the italian association for artificial intelligence, genova, italy, november 29–December 1, 2016, proceedings* Springer.

Afridi, M. J., Ross, A., & Shapiro, E. M. (2017). On automated source selection for transfer learning in convolutional neural networks. *Pattern Recognition,*

Agrawal, A., & Mishra, N. K. (2016). Fusion based emotion recognition system. 727-732.

Ahram, T., & Falcão, C.Advances in human factors in wearable technologies and game design.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine, 35*(4), 105-120.

Arias, J. P., Busso, C., & Yoma, N. B. (2014). Shape-based modeling of the fundamental frequency contour for emotion detection in speech. *Computer Speech & Language, 28*(1), 278-294.

Bai, X., Shi, B., Zhang, C., Cai, X., & Qi, L. (2017). Text/non-text image classification in the wild with convolutional neural networks. *Pattern Recognition, 66*, 437-446.

Bai, X., Yang, M., Lyu, P., & Xu, Y. (2017). Integrating scene text and visual appearance for fine-grained image classification with convolutional neural networks. *arXiv Preprint arXiv:1704.04613,*

Barros, P., Jirak, D., Weber, C., & Wermter, S. (2015). Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks, 72*, 140-151.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. *Computer vision–ECCV 2006,* , 404-417.

Bell, S., Bala, K., & Snavely, N. (2014). Intrinsic images in the wild. *ACM Transactions on Graphics (TOG), 33*(4), 159.

Bertero, D., & Fung, P. (2017). A first look into a convolutional neural network for speech emotion detection. 5115-5119.

Bhatia, S. K., Mishra, K. K., Tiwari, S., & Singh, V. K. (2016a). Advances in computer and computational sciences. *Proceedings of ICCCCS, 1*

Bhatia, S. K., Mishra, K. K., Tiwari, S., & Singh, V. K. (2016b). Advances in computer and computational sciences. *Proceedings of ICCCCS, 1*

Bhattacharyya, S. S., van, d. S., Atan, O., Tekin, C., & Sudusinghe, K. (2014). Data-driven stream mining systems for computer vision. *Advances in embedded computer vision* (pp. 249-264) Springer.

Bowyer, K. W., & Burge, M. J. (2016). *Handbook of iris recognition* Springer.

Budil, D. E., Lee, S., Saxena, S., & Freed, J. H. (1996). Nonlinear-least-squares analysis of slow-motion EPR spectra in one and two dimensions using a modified Levenberg–Marquardt algorithm. *Journal of Magnetic Resonance, Series A, 120*(2), 155-189.

Burlina, P., Pacheco, K. D., Joshi, N., Freund, D. E., & Bressler, N. M. (2017). Comparing humans and deep learning performance for grading AMD: A study in using universal deep features and transfer learning for automated AMD analysis. *Computers in Biology and Medicine, 82*, 80-86.

Callaway, J., & Rozar, T. (2015). Quantified wellness wearable technology usage and market summary. *RGA Reinsurance Company,*

Camps-Valls, G., Bioucas-Dias, J., & Crawford, M. (2016). A special issue on advances in machine learning for remote sensing and geosciences from the guest editors]. *IEEE Geoscience and Remote Sensing Magazine, 4*(2), 5-7.

Camps-Valls, G., Tuia, D., Bruzzone, L., & Benediktsson, J. A. (2014). Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE Signal Processing Magazine, 31*(1), 45-54.

Chen, H., Tsai, S. S., Schroth, G., Chen, D. M., Grzeszczuk, R., & Girod, B. (2011). Robust text detection in natural images with edge-enhanced maximally stable extremal regions. 2609-2612.

Cherian, A., & Sra, S. (2016). Positive definite matrices: Data representation and applications to computer vision. *Algorithmic Advances in Riemannian Geometry and Applications: For Machine Learning, Computer Vision, Statistics, and Optimization,* , 93.

Chernova, S., & Thomaz, A. L. (2014). Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 8*(3), 1-121.

ChunRong, C., ShanXiong, C., Lin, C., & YuChen, Z. (2017). Method for solving LASSO problem based on multidimensional weight. *Advances in Artificial Intelligence, 2017*

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. *, 1*(1-22) 1-2.

Cui, Y., Zhou, F., Lin, Y., & Belongie, S. (2016). Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. 1153-1162.

Dandois, J. P., & Ellis, E. C. (2013). High spatial resolution three-dimensional mapping of vegetation spectral dynamics using computer vision. *Remote Sensing of Environment, 136*, 259-276.

Dandois, J. P., Olano, M., & Ellis, E. C. (2015). Optimal altitude, overlap, and weather conditions for computer vision UAV estimates of forest structure. *Remote Sensing, 7*(10), 13895-13920.

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research best of the web]. *IEEE Signal Processing Magazine, 29*(6), 141-142.

Devin, C., Gupta, A., Darrell, T., Abbeel, P., & Levine, S. (2017). Learning modular neural network policies for multi-task and multi-robot transfer. 2169-2176.

Ding, Z., & Fu, Y. (2017). Robust transfer metric learning for image classification. *IEEE Transactions on Image Processing, 26*(2), 660-670.

Dixit, A., Pal, A. K., Temghare, S., & Mapari, V. (2017). Emotion detection using decision tree. *Development, 4*(2)

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. 647-655.

Douiji, Y., & Mousanif, H. (2015). I-CARE: Intelligent context aware system for recognizing emotions from text. 1-5.

Egan, D., Brennan, S., Barrett, J., Qiao, Y., Timmerer, C., & Murray, N. (2016). An evaluation of heart rate and ElectroDermal activity as an objective QoE evaluation method for immersive virtual reality environments. 1-6.

Endler, M., Briot, J., Almeida, V. P., Silva, F. S. E., & Haeusler, E. (2017). Towards stream-based reasoning and machine learning for IoT applications.

Fedor, S., Chau, P., Bruno, N., Picard, R. W., Camprodon, J., & Hale, T. (2016). Can we predict depression from the asymmetry of electrodermal activity? *Journal of Medical Internet Research, 18*(12)

Filko, D., & Martinović, G. (2013). Emotion recognition system by a neural network based facial expression analysis. *Automatika, 54*(2), 263-272.

Friedrich, G., Helmert, M., & Wotawa, F. (2016). *KI 2016: Advances in artificial intelligence: 39th annual german conference on AI, klagenfurt, austria, september 26-30, 2016, proceedings*Springer.

Fsr, A., & Torresen, J. (2016). Smartphone accelerometer data used for detecting human emotions. 410-415.

Gall, J., & Lempitsky, V. (2013). Class-specific hough forests for object detection. *Decision forests for computer vision and medical image analysis* (pp. 143-157) Springer.

Gao, J., Ling, H., Hu, W., & Xing, J. (2014a). Transfer learning based visual tracking with gaussian processes regression. 188-203.

Gao, J., Ling, H., Hu, W., & Xing, J. (2014b). Transfer learning based visual tracking with gaussian processes regression. 188-203.

Garbarino, M., Lai, M., Bender, D., Picard, R. W., & Tognetti, S. (2014). Empatica E3—A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. 39-42.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2015). *The KITTI vision benchmark suite*

Geman, D., Geman, S., Hallonquist, N., & Younes, L. (2015). Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences, 112*(12), 3618-3623.

Ghamisi, P., Souza, R., Benediktsson, J. A., Zhu, X. X., Rittner, L., & Lotufo, R. A. (2016). Extinction profiles for the classification of remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing, 54*(10), 5631-5645.

Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., & Provost, E. M. (2017). Progressive neural networks for transfer learning in emotion recognition. *arXiv Preprint arXiv:1706.03256,*

Gonzalez, A., Bergasa, L. M., Yebes, J. J., & Bronte, S. (2012). Text location in complex images. 617-620.

Goodwin, M. S. (2016). 28.2 laboratory and home-based assessment of electrodermal activity in individuals with autism spectrum disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 55*(10), S302.

Greco, A., Valenza, G., Lanata, A., Rota, G., & Scilingo, E. P. (2014). Electrodermal activity in bipolar patients during affective elicitation. *IEEE Journal of Biomedical and Health Informatics, 18*(6), 1865-1873.

Grgic, M., & Delac, K. (2013). Face recognition homepage. *Zagreb, Croatia (Www.Face-Rec.Org/Databases), 324*

Guarino, N., & Guizzardi, G. (2016). Relationships and events: Towards a general theory of reification and truthmaking. *AI* IA 2016 advances in artificial intelligence* (pp. 237-249) Springer.

Guillemot, C., & Le Meur, O. (2014a). Image inpainting: Overview and recent advances. *IEEE Signal Processing Magazine, 31*(1), 127-144.

Guillemot, C., & Le Meur, O. (2014b). Image inpainting: Overview and recent advances. *IEEE Signal Processing Magazine, 31*(1), 127-144.

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion recognition using bio-sensors: First steps towards an automatic system. 36-48.

Hager, G. D., Bryant, R., Horvitz, E., Mataric, M., & Honavar, V. (2017). Advances in artificial intelligence require progress across all of computer science. *arXiv Preprint arXiv:1707.04352,*

Han, J., Shao, L., Xu, D., & Shotton, J. (2013). Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics, 43*(5), 1318-1334.

Hasan, H., & Abdul-Kareem, S. (2014). Human–computer interaction using vision-based hand gesture recognition systems: A survey. *Neural Computing and Applications, 25*(2), 251-261.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245-258.

He, C., Yao, Y., & Ye, X. (2017). An emotion recognition system based on physiological signals obtained by wearable sensors. *Wearable sensors and robots* (pp. 15-25) Springer.

He, T., Huang, W., Qiao, Y., & Yao, J. (2016). Text-attentional convolutional neural network for scene text detection. *IEEE Transactions on Image Processing, 25*(6), 2529-2541.

Hewig, J., Hagemann, D., Seifert, J., Gollwitzer, M., Naumann, E., & Bartussek, D. (2005). Brief report. *Cognition & Emotion, 19*(7), 1095-1109.

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics, 3*(2), 119-131.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, 79*(8), 2554-2558.

Ingle, B. L., Veber, B. C., Nichols, J. W., & Tornero-Velez, R. (2016). Informing the human plasma protein binding of environmental chemicals by machine learning in the pharmaceutical space: Applicability domain and limits of predictability. *Journal of Chemical Information and Modeling, 56*(11), 2243-2252.

Ioannidou, A., Chatzilari, E., Nikolopoulos, S., & Kompatsiaris, I. (2017). Deep learning advances in computer vision with 3D data: A survey. *ACM Computing Surveys (CSUR), 50*(2), 20.

Iqbal, K., Yin, X., Hao, H., Asghar, S., & Ali, H. (2014). Bayesian network scores based text localization in scene images. 2218-2225.

Iqbal, K., Yin, X., Yin, X., Ali, H., & Hao, H. (2013). Classifier comparison for MSER-based text classification in scene images. 1-6.

Iscen, A., Tolias, G., Gosselin, P., & Jégou, H. (2015). A comparison of dense region detectors for image search and fine-grained classification. *IEEE Transactions on Image Processing, 24*(8), 2369-2381.

Izumi, M., Manabe, E., Uematsu, S., Watanabe, A., & Moritani, T. (2016). Changes in autonomic nervous system activity, body weight, and percentage fat mass in the first year postpartum and factors regulating the return to pre-pregnancy weight. *Journal of Physiological Anthropology, 35*(1), 26.

Jain, A., Wojcik, B., Joachims, T., & Saxena, A. (2013). Learning trajectory preferences for manipulators via iterative improvement. 575-583.

Janoch, A., Karayev, S., Jia, Y., Barron, J. T., Fritz, M., Saenko, K., & Darrell, T. (2013). A category-level 3d object dataset: Putting the kinect to work. *Consumer depth cameras for computer vision* (pp. 141-165) Springer.

Jiang, R., Ho, A. T. S., Cheheb, I., Al-Maadeed, N., Al-Maadeed, S., & Bouridane, A. (2017). Emotion recognition from scrambled facial images via many graph embedding. *Pattern Recognition, 67*, 245-251.

Jiang, Y., Koppula, H., & Saxena, A. (2013). Hallucinated humans as the hidden context for labeling 3d scenes. 2993-3000.

Kang, L., Li, Y., & Doermann, D. (2014). Orientation robust text line detection in natural images. 4034-4041.

Karaoglu, S., Tao, R., Gevers, T., & Smeulders, A. W. M. (2017). Words matter: Scene text for image classification and retrieval. *IEEE Transactions on Multimedia, 19*(5), 1063-1076.

Kelsey, M., Palumbo, R. V., Urbaneja, A., Akcakaya, M., Huang, J., Kleckner, I. R., . . . Goodwin, M. S. (2017). Artifact detection in electrodermal activity using sparse recovery. *, 10211* 1.

Keltner, D., & CORDARO, D. T. (2017). Understanding multimodal emotional expressions. *The Science of Facial Expression,* , 1798.

Kensinger, E. A., & Gutchess, A. H. (2017). Cognitive aging in a social and affective context: Advances over the past 50 years. *The Journals of Gerontology: Series B, 72*(1), 61-70.

Khoury, R., & Drummond, C. (2016). *Advances in artificial intelligence: 29th canadian conference on artificial intelligence, canadian AI 2016, victoria, BC, canada, may 31-june 3, 2016. proceedings* Springer.

Klette, R. (2014). *Concise computer vision* Springer.

Koch, C., Georgieva, K., Kasireddy, V., Akinci, B., & Fieguth, P. (2015a). A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. *Advanced Engineering Informatics, 29*(2), 196-210.

Koch, C., Georgieva, K., Kasireddy, V., Akinci, B., & Fieguth, P. (2015b). A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. *Advanced Engineering Informatics, 29*(2), 196-210.

Kolosnjaji, B., Zarras, A., Webster, G., & Eckert, C. (2016). Deep learning for classification of malware system call sequences. 137-149.

Koppula, H. S., Gupta, R., & Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research, 32*(8), 951-970.

Kovashka, A., Russakovsky, O., Fei-Fei, L., & Grauman, K. (2016). Crowdsourcing in computer vision. *Foundations and Trends® in Computer Graphics and Vision, 10*(3), 177-243.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. 1097-1105.

Kulis, B. (2013a). Metric learning: A survey. *Foundations and Trends® in Machine Learning, 5*(4), 287-364.

Kulis, B. (2013b). Metric learning: A survey. *Foundations and Trends® in Machine Learning, 5*(4), 287-364.

Kumar, S., Gao, X., & Welch, I. (2016). Learning under data shift for domain adaptation: A model-based co-clustering transfer learning solution. 43-54.

Kumar, V. A., Gupta, S., Chandra, S. S., Raman, S., & Channappayya, S. S. (2017). No-reference quality assessment of tone mapped high dynamic range (HDR) images using transfer learning. 1-3.

Kupcsik, A., Hsu, D., & Lee, W. S. (2018a). Learning dynamic robot-to-human object handover from human feedback. *Robotics research* (pp. 161-176) Springer.

Kupcsik, A., Hsu, D., & Lee, W. S. (2018b). Learning dynamic robot-to-human object handover from human feedback. *Robotics research* (pp. 161-176) Springer.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). International affective picture system (IAPS): Technical manual and affective ratings. *Gainesville, FL: The Center for Research in Psychophysiology, University of Florida, 2*

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.

Lee, M., Bressler, S., & Kozma, R. (2017). *Advances in cognitive engineering using neural networks* Elsevier.

Lei, J., Rao, Y., Li, Q., Quan, X., & Wenyin, L. (2014). Towards building a social emotion detection system for online news. *Future Generation Computer Systems, 37*, 438-448.

Lera, G., & Pinzolas, M. (2002). Neighborhood based levenberg-marquardt algorithm for neural network training. *IEEE Transactions on Neural Networks, 13*(5), 1200-1203.

Li, Z., Su, C., Li, G., & Su, H. (2015). Fuzzy approximation-based adaptive backstepping control of an exoskeleton for human upper limbs. *IEEE Transactions on Fuzzy Systems, 23*(3), 555-566.

Likforman-Sulem, L., Esposito, A., Faundez-Zanuy, M., Clémençon, S., & Cordasco, G. (2017). EMOTHAW: A novel database for emotional state recognition from handwriting and drawing.*IEEE Transactions on Human-Machine Systems, 47*(2), 273-284.

Liu, J., Wang, S., Turkbey, E. B., Linguraru, M. G., Yao, J., & Summers, R. M. (2015). Computer-aided detection of renal calculi from noncontrast CT images using TV-flow and MSER features. *Medical Physics, 42*(1), 144-153.

Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., & Chen, X. (2014). Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. 494-501.

Liu, Y., Zhang, Y., Zhang, X., & Liu, C. (2016). Adaptive spatial pooling for image classification. *Pattern Recognition, 55*, 58-67.

Loconsole, C., Miranda, C. R., Augusto, G., Frisoli, A., & Orvalho, V. (2014). Real-time emotion recognition novel method for geometrical facial features extraction. *, 1* 378-385.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. 3431-3440.

Long, M., Wang, J., Ding, G., Pan, S. J., & Philip, S. Y. (2014a). Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 26*(5), 1076-1089.

Long, M., Wang, J., Ding, G., Pan, S. J., & Philip, S. Y. (2014b). Adaptation regularization: A general framework for transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 26*(5), 1076-1089.

Long, M., Wang, J., Ding, G., Shen, D., & Yang, Q. (2014a). Transfer learning with graph co-regularization. *IEEE Transactions on Knowledge and Data Engineering, 26*(7), 1805-1818.

Long, M., Wang, J., Ding, G., Shen, D., & Yang, Q. (2014b). Transfer learning with graph co-regularization. *IEEE Transactions on Knowledge and Data Engineering, 26*(7), 1805-1818.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *, 2* 1150-1157.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110.

Lu, R., Li, Z., Su, C., & Xue, A. (2014). Development and learning control of a human limb with a rehabilitation exoskeleton. *IEEE Transactions on Industrial Electronics, 61*(7), 3776-3785.

Ludwig, K., Ross, E., Langhals, N., Weber, D., Luis Lujan, J., & Georgakopoulos, D. (2017). The autonomic nervous system. *NEUROPROSTHETICS: Theory and practice* (pp. 12-39) World Scientific.

Lv, Z., Halawani, A., Feng, S., Ur Réhman, S., & Li, H. (2015). Touch-less interactive augmented reality game on vision-based wearable device. *Personal and Ubiquitous Computing, 19*(3-4), 551-567.

Ma, J., Sun, D., Qu, J., Liu, D., Pu, H., Gao, W., & Zeng, X. (2016). Applications of computer vision for assessing quality of agri-food products: A review of recent research advances. *Critical Reviews in Food Science and Nutrition, 56*(1), 113-127.

Ma, Y., & Guo, G. (2014a). *Support vector machines applications* Springer Science & Business Media.

Ma, Y., & Guo, G. (2014b). *Support vector machines applications* Springer Science & Business Media.

Maaoui, C., & Pruski, A. (2010). Emotion recognition through physiological signals for human-machine communication. *Cutting edge robotics 2010* () InTech.

Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based human–computer interaction. *Advances in physiological computing* (pp. 39-65) Springer.

Martinel, N., Prati, A., & Micheloni, C. (2014). Distributed mobile computer vision: Advances, challenges and applications. *Distributed embedded smart cameras* (pp. 93-120) Springer.

Martinez, B., & Valstar, M. F. (2016). Advances, challenges, and opportunities in automatic facial expression recognition. *Advances in face detection and facial image analysis* (pp. 63-100) Springer.

Marzuki, A., Rumpa, L. D., Wibawa, A. D., & Purnomo, M. H. (2016). Classification of human state emotion from physiological signal pattern using pulse sensor based on learning vector quantization. 129-134.

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing, 22*(10), 761-767.

Matuszek, C., Bo, L., Zettlemoyer, L., & Fox, D. (2014a). Learning from unscripted deictic gesture and language for human-robot interactions. 2556-2563.

Matuszek, C., Bo, L., Zettlemoyer, L., & Fox, D. (2014b). Learning from unscripted deictic gesture and language for human-robot interactions. 2556-2563.

McDonnell, D. (2014). *Electrodermal activity sensor* Google Patents.

Michalopoulos, K., & Bourbakis, N. (2017). Application of multiscale entropy on EEG signals for emotion detection. 341-344.

Minh, H. Q., & Murino, V. (2016). *Algorithmic advances in riemannian geometry and applications: For machine learning, computer vision, statistics, and optimization* Springer.

Mitchell, T. M., Cohen, W. W., Hruschka Jr, E.,R., Talukdar, P. P., Betteridge, J., Carlson, A., . . . Krishnamurthy, J. (2015). Never ending learning. 2302-2310.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . Ostrovski, G. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529-533.

Moré, J.,J. (1978). The levenberg-marquardt algorithm: Implementation and theory. *Numerical analysis* (pp. 105-116) Springer.

Morrison, M. A., Zimmerman, M. W., Look, A. T., & Stewart, R. A. (2016). Studying the peripheral sympathetic nervous system and neuroblastoma in zebrafish. *Methods in Cell Biology, 134*, 97-138.

Mosavi, A., & Varkonyi, A. (2016). *Integration of machine learning and optimization for robot learning, advances in intelligent systems and computing* Springer-Verlag Berlin Heidelberg.

Mozafari, B., Sarkar, P., Franklin, M., Jordan, M., & Madden, S. (2014a). Scaling up crowd-sourcing to very large datasets: A case for active learning. *Proceedings of the VLDB Endowment, 8*(2), 125-136.

Mozafari, B., Sarkar, P., Franklin, M., Jordan, M., & Madden, S. (2014b). Scaling up crowd-sourcing to very large datasets: A case for active learning. *Proceedings of the VLDB Endowment, 8*(2), 125-136.

Nam, J., Pan, S. J., & Kim, S. (2013). Transfer defect learning. 382-391.

Neji, M., Ammar, M. B., Wali, A., & Alimi, A. M. (2013). Towards an intelligent information research system based on the human behavior: Recognition of user emotional state. 371-376.

Neumann, L., & Matas, J. (2012). Real-time scene text localization and recognition. 3538-3545.

Neves, A. J. R., Trifan, A., Cunha, B., & Azevedo, J. L. (2016). Real-time color coded object detection using a modular computer vision library. *Advances in Computer Science: An International Journal, 5*(1), 110-123.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016a). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. 3387-3395.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016b). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. 3387-3395.

Nguyen, B. T., Trinh, M. H., Phan, T. V., & Nguyen, H. D. (2017). An efficient real-time emotion detection using camera and facial landmarks. 251-255.

Nikolaidis, S., Ramakrishnan, R., Gu, K., & Shah, J. (2015). Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. 189-196.

Nisa'Minhad, K., Ali, S. H. M., Khai, J. O. S., & Ahmad, S. A. (2016). Human emotion classifications for automotive driver using skin conductance response signal. 371-375.

Noh, H., Hongsuck Seo, P., & Han, B. (2016). Image question answering using convolutional neural network with dynamic parameter prediction. 30-38.

Norman, D. (2017). Design, business models, and human-technology teamwork: As automation and artificial intelligence technologies develop, we need to think less about human-machine interfaces and more about human-machine teamwork. *Research-Technology Management, 60*(1), 26-30.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal of Medicine, 375*(13), 1216.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. 1717-1724.

Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2015a). Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv Preprint arXiv:1511.06342,*

Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2015b). Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv Preprint arXiv:1511.06342,*

Patel, V. M., Gopalan, R., Li, R., & Chellappa, R. (2015). Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine, 32*(3), 53-69.

Peng, P., Tian, Y., Xiang, T., Wang, Y., Pontil, M., & Huang, T. (2017). Joint semantic and latent attribute modelling for cross-class transfer learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*

Picard, R. W. (2015). Recognizing stress, engagement, and positive emotion. 3-4.

Picard, R. W., Fedor, S., & Ayzenberg, Y. (2016). Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion Review, 8*(1), 62-75.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(10), 1175-1191.

Pirsiavash, H., Vondrick, C., & Torralba, A. (2014a). Assessing the quality of actions. 556-571.

Pirsiavash, H., Vondrick, C., & Torralba, A. (2014b). Assessing the quality of actions. 556-571.

Pogue, J. R., Cloutier, R. M., Russo, M. J., McKinnis, S. A., & Blumenthal, H. (2016). Electrodermal activity and anxiety symptoms among adolescent females. *, 53* S40.

Pool, C., & Nissim, M. (2016). Distant supervision for emotion detection using facebook reactions. *arXiv Preprint arXiv:1611.02988,*

Poria, S., Cambria, E., Gelbukh, A., Bisio, F., & Hussain, A. (2015). Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine, 10*(4), 26-36.

Povinelli, R. J., Johnson, M. T., Lindgren, A. C., & Ye, J. (2004). Time series classification using gaussian mixture models of reconstructed phase spaces. *IEEE Transactions on Knowledge and Data Engineering, 16*(6), 779-783.

Prahm, C., Paassen, B., Schulz, A., Hammer, B., & Aszmann, O. (2017). Transfer learning for rapid re-calibration of a myoelectric prosthesis after electrode shift. *Converging clinical and engineering research on neurorehabilitation II* (pp. 153-157) Springer.

Prince, E. B., Kim, E. S., Wall, C. A., Gisin, E., Goodwin, M. S., Simmons, E. S., . . . Shic, F. (2017). The relationship between autism symptoms and arousal level in toddlers with autism spectrum disorder, as measured by electrodermal activity. *Autism, 21*(4), 504-508.

Qu, Y., Wu, S., Liu, H., Xie, Y., & Wang, H. (2014). Evaluation of local features and classifiers in BOW model for image classification. *Multimedia Tools and Applications, 70*(2), 605-624.

Ragot, M., Martin, N., Em, S., Pallamin, N., & Diverrez, J. (2017). Emotion recognition using physiological signals: Laboratory vs. wearable sensors. 15-22.

Raja, M., & Sigg, S. (2016). Applicability of rf-based methods for emotion recognition: A survey. 1-6.

Rajesh, K. M., & Naveenkumar, M. (2016). A robust method for face recognition and face emotion detection system using support vector machines. 1-5.

Rakshit, R., Reddy, V. R., & Deshpande, P. (2016). Emotion detection and recognition using HRV features derived from photoplethysmogram signals. 2.

Rao, Y., Li, Q., Wenyin, L., Wu, Q., & Quan, X. (2014). Affective topic model for social emotion detection. *Neural Networks, 58*, 29-37.

Roggero, E., Pérez, A. R., Pollachini, N., Villar, S. R., Wildmann, J., Besedovsky, H., & del Rey, A. (2016). The sympathetic nervous system affects the susceptibility and course of trypanosoma cruzi infection. *Brain, Behavior, and Immunity, 58*, 228-236.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1*, 45-76.

Sampson, P., Freeman, C., Coote, S., Demain, S., Feys, P., Meadmore, K., & Hughes, A. (2016). Using functional electrical stimulation mediated by iterative learning control and robotics to improve arm movement for people with multiple sclerosis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 24*(2), 235-248.

Schwartz, E. L. (1980). Computational anatomy and functional architecture of striate cortex: A spatial mapping approach to perceptual coding. *Vision Research, 20*(8), 645-669.

Semwal, N., Kumar, A., & Narayanan, S. (2017). Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models. 1-6.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. 806-813.

Shi, Z., Siva, P., & Xiang, T. (2017a). Transfer learning by ranking for weakly supervised object annotation. *arXiv Preprint arXiv:1705.00873,*

Shi, Z., Siva, P., & Xiang, T. (2017b). Transfer learning by ranking for weakly supervised object annotation. *arXiv Preprint arXiv:1705.00873,*

Shin, H., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., . . . Summers, R. M. (2016a). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging, 35*(5), 1285-1298.

Shin, H., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., . . . Summers, R. M. (2016b). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging, 35*(5), 1285-1298.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint arXiv:1409.1556,*

Sing, D. C., Metz, L. N., & Dudli, S. (2017). Machine learning-based classification of 38 years of spine-related literature into 100 research topics. *Spine, 42*(11), 863-870.

Singh, T. P., Chatli, M. K., Singh, P., & Kumar, P. (2013). Advances in computer vision technology for foods of animal and aquatic origin (a). *J Meat Sci Technol, 1*, 40-49.

Smisek, J., Jancosek, M., & Pajdla, T. (2013a). 3D with kinect. *Consumer depth cameras for computer vision* (pp. 3-25) Springer.

Smisek, J., Jancosek, M., & Pajdla, T. (2013b). 3D with kinect. *Consumer depth cameras for computer vision* (pp. 3-25) Springer.

Socher, R. (2014a). *Recursive deep learning for natural language processing and computer vision* Citeseer.

Socher, R. (2014b). *Recursive deep learning for natural language processing and computer vision* Citeseer.

Soleymani, M., Asghari-Esfeden, S., Fu, Y., & Pantic, M. (2016). Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing, 7*(1), 17-28.

Soleymani, M., Asghari-Esfeden, S., Pantic, M., & Fu, Y. (2014). Continuous emotion detection using EEG signals and facial expressions. 1-6.

Sonka, M., Hlavac, V., & Boyle, R. (2014). *Image processing, analysis, and machine vision* Cengage Learning.

Sun, D. (2016). *Computer vision technology for food quality evaluation* Academic Press.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. 2818-2826.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309-318.

Teutsch, M., Muller, T., Huber, M., & Beyerer, J. (2014). Low resolution person detection with a moving thermal infrared camera by hot spot classification. 209-216.

Tiple, B., & Thomas, P. A. (2017). Analysis of features for mood detection in north indian classical music-A literature review. *Ijrcct, 6*(6), 181-185.

Uhr, L. (2014a). *Parallel computer vision* Elsevier.

Uhr, L. (2014b). *Parallel computer vision* Elsevier.

Umbaugh, S. E. (2016). *Digital image processing and analysis: Human and computer vision applications with CVIPtools* CRC press.

Valenza, G., & Scilingo, E. P. (2014). Conclusions and discussion on mood and emotional-state recognition using the autonomic nervous system dynamics. *Autonomic nervous system dynamics for mood and emotional-state recognition* (pp. 127-138) Springer.

Veenendaal, A., Daly, E., Jones, E., Gang, Z., Vartak, S., & Patwardhan, R. S. (2014). Group emotion detection using edge detecttion mesh analysis. *Computer Science and Emerging Research Journal, 2*

Wang, X., Nie, D., & Lu, B. (2014). Emotional state classification from EEG data using machine learning approach. *Neurocomputing, 129*, 94-106.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* Morgan Kaufmann.

Xie, L., Tian, Q., Wang, M., & Zhang, B. (2014). Spatial pooling of heterogeneous features for image classification. *IEEE Transactions on Image Processing, 23*(5), 1994-2008.

Xie, L., Wang, J., Zhang, B., & Tian, Q. (2016). Incorporating visual adjectives for image classification. *Neurocomputing, 182*, 48-55.

Xu, Y., Hübener, I., Seipp, A., Ohly, S., & David, K. (2017). From the lab to the real-world: An investigation on the influence of human movement on emotion recognition using physiological signals. 345-350.

y Gómez, M. M., Escalante, H. J., Segura, A., & de, D. M. (2016). *Advances in artificial intelligence-IBERAMIA 2016: 15th ibero-american conference on AI, san josé, costa rica, november 23-25, 2016, proceedings* Springer.

Yang, Y., Luo, H., Xu, H., & Wu, F. (2016). Towards real-time traffic sign detection and classification. *IEEE Transactions on Intelligent Transportation Systems, 17*(7), 2022-2031.

Yang, Z., Salakhutdinov, R., & Cohen, W. W. (2017). Transfer learning for sequence tagging with hierarchical recurrent networks. *arXiv Preprint arXiv:1703.06345,*

Ye, G., & Alterovitz, R. (2017). Guided motion planning. *Robotics research* (pp. 291-307) Springer.

Ye, J., Povinelli, R. J., & Johnson, M. T. (2002). Phoneme classification using naive bayes classifier in reconstructed phase space. 37-40.

Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., & Gall, J. (2013a). A survey on human motion analysis from depth data. *Time-of-flight and depth imaging. sensors, algorithms, and applications* (pp. 149-187) Springer.

Ye, M., Zhang, Q., Wang, L., Zhu, J., Yang, R., & Gall, J. (2013b). A survey on human motion analysis from depth data. *Time-of-flight and depth imaging. sensors, algorithms, and applications* (pp. 149-187) Springer.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? 3320-3328.

Yu, B., & Wan, H. (2016). Chinese text detection and recognition in natural scene using HOG and SVM. *DEStech Transactions on Computer Science and Engineering,*

Yu, F., Chang, E., Xu, Y., & Shum, H. (2001). Emotion detection from speech to enrich multimedia content. *Advances in Multimedia Information processing—PCM 2001,* , 550-557.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv Preprint arXiv:1506.03365,*

Zadeh, L. A. (2016). *Method and system for analyzing and recognition of an emotion or expression from multimedia, text, or sound track* Google Patents.

Zhang, B., Huang, W., Gong, L., Li, J., Zhao, C., Liu, C., & Huang, D. (2015). Computer vision detection of defective apples using automatic lightness correction and weighted RVM classifier. *Journal of Food Engineering, 146*, 143-151.

Zhang, B., Huang, W., Li, J., Zhao, C., Fan, S., Wu, J., & Liu, C. (2014). Principles, developments and applications of computer vision for external quality inspection of fruits and vegetables: A review. *Food Research International, 62*, 326-343.

Zhang, Z., Cui, L., Liu, X., & Zhu, T. (2016). Emotion detection using kinect 3D facial points. 407-410.

Zhao, M., Adib, F., & Katabi, D. (2016). Emotion recognition using wireless signals. 95-108.