

A Machine Learning approach to Febrile Classification

by

Theodore Paul Kostopoulos

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

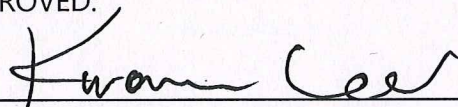
Degree of Master of Science

in

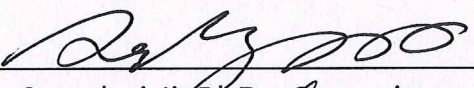
Biomedical Engineering

APPROVED:

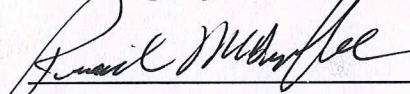
May 2018



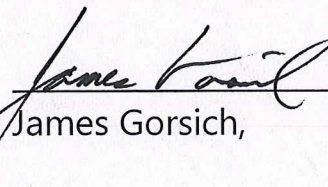
Dr. Kwonmoo Lee, PhD., Major Advisor



Dr. Songbaj Ji, PhD., Committee Member



Richard McDuffie, Committee Member



James Gorsich, Committee Member

Acknowledgement

I would like to thank my family for all their love and support while conducting this research

I would also like to thank Chauncey Wang and the Quantitative Cellular Imaging Laboratory for their mentorship and technical support in Machine Learning during each phase of the research.

I would like to thank Helen of Troy for the acquisition of the thermograms used in this research.

Abstract

General health screening is needed to decrease the risk of pandemic in high volume areas. Thermal characterization, via infrared imaging, is an effective technique for fever detection, however, strict use requirements in combination with highly controlled environmental conditions compromise the practicality of such a system. Combining advanced processing techniques to thermograms of individuals can remove some of these requirements allowing for more flexible classification algorithms. The purpose of this research was to identify individuals who had febrile status utilizing modern thermal imaging and machine learning techniques in a minimally controlled setting. Two methods were evaluated with data that contained environmental, and acclimation noise due to data gathering technique. The first was a pretrained VGG16 Convolutional Neural Network found to have F1 score of 0.77 (accuracy of 76%) on a balanced dataset. The second was a VGG16 Feature Extractor that gives inputs to a principle components analysis and utilizes a support vector machine for classification. This technique obtained a F1 score of 0.84 (accuracy of 85%) on balanced data sets. These results demonstrate that machine learning is an extremely viable technique to classify febrile status independent of noise affiliated.

Table of Contents

Table of Equations	viii
1. Introduction	1
1.1. Motivation	1
1.2. Thermodynamics and Physiological Heat Transfer	5
1.3. Human Temperature Measurements	8
1.4. Radiation Based Measurements	11
1.5. Infrared Measurements in Practice	13
1.6. Brief background of Machine Learning	20
1.7. Deep Learning Functionality	21
2. Data Gathered and Equipment	27
2.1. Clinical Data	27
2.2. Thermal Imager	28
2.3. Computer Specifications	28
2.4. GPU	29
3. Research Overview	30
Hypothesis	30

4.	Localized Area Investigation	35
4.1.	Purpose	35
4.2.	Results	38
5.	Binary Classification with Pretrained VGG16	43
5.1.	Implementation	43
5.2.	Experiment Results	46
5.3.	Conclusion	49
6.	VGG16 Feature Extractor to a Principle Components Analysis to Support Vector Machine Approach	50
6.1.	Execution	50
6.2.	Results	56
7.	Final Experiment: Comparison the Two Methods	59
7.1.	Design	60
7.2.	Results	60
7.3.	Conclusion	61
8.	Discussion	62
8.1.	Primary Scope	62

A.....Bibliography
..... A
B.....Glossary of Terms
..... E

Table of Figures

Figure 1 Number of Influenza Cases in the Population.....	2
Figure 2 <i>Prevalence</i> of the Flu in 1918 and 2009	3
Figure 3 World Population verse Flu Cases 1918 and 2009.....	4
Figure 4 Electromagnetic Spectrum [48].....	11
Figure 5 Spectral responses for different HgCdTe alloy detectors [35]	12
Figure 6 Welch Allyn Pro6000 Tympanic Thermometer.....	14
Figure 7 Exergen Temporal Thermometer Use [53].....	15
Figure 8 Braun NTF3000 [36].....	15
Figure 9 Visual Demonstration of Overfit.....	24
Figure 10 Focus of Localized Area Investigation.....	35
Figure 11 Extraction of Sites	37
Figure 12 Squared Error verse Power Tested	39
Figure 13 Bland Altman Plot of Best Sites	42
Figure 14 Updated Approach with Pretrained Model.....	43
Figure 15 Structured Crop Example VGG16.....	45
<i>Figure 16 VGG16 Architecture</i>	46
Figure 17 Slope Graph Comparing the Base Pretrained VGG16 Network with the Semi-Randomly Down Sampled VGG16 Network.....	48
Figure 18 Final Approach for Classification.....	50
Figure 19 Structure Crop with Increased Contrast.....	51
Figure 20 Base image (left) and Image raised to 10th power(right).....	52
Figure 21 Evaluation for Point by Point Power.....	53
Figure 22 Circular Distribution versus Elliptical Distribution with vector.....	54
Figure 23 Optimal Separating Hyperplane [14].....	55
Figure 24 Accuracy Heatmap of SVM-PCA Approach	56
Figure 25 PCA-SVM-Vote! Accuracy Heatmap	58
Figure 26 Outliers Identified in PCA-SVM-Vote! Results	58
Figure 27 Comparison of the Pretrained VGG16 with the PCA-SVM-Vote! Algorithm.....	61
Figure 28 Unacclimated Patient.....	63

Table of Tables

Table 1 Bitar, Goubar and Desenclos Research Comparison [3]	18
Table 2 FLIR T660 Specifications	28
Table 3 CPU Specifications	29
Table 4 GPU Specifications	29
Table 5 Equations Outputted by Site	40
Table 6 Output of Least Square Sum Regression by Site	40
Table 7 Further Results Canthus and Temple Combined	41
Table 8 Pretrained VGG16 Results with the Full Face and Structured Crop	46
Table 9 Results of the First Down Sampled Network	48
Table 10 PCA-SVM Results	56
Table 11 Feature Extractor - PCA - SVM - Vote Average Results	57

Table of Equations

Equation 1 ReLU Activation Function.....	21
Equation 2 Softmax Activation Function.....	22
Equation 3 Binary Cross Entropy.....	23
Equation 4 F1 Score.....	30
Equation 5 Modified External Heat Transfer Equation.....	38
Equation 6 Weighted Binary Cross Entropy Equation.....	47

1. Introduction

1.1. Motivation

The presence of disease is lurking behind every human interaction. Disease spreads in various modalities, such as contact, contaminated liquids and airborne pathogens. The probability of an individual becoming ill increases with prolonged exposure to others already afflicted with a given disease. Pandemics and epidemics typically arise when a foreign pathogen is introduced to a population, and the individuals are subject to increased exposure to this pathogen. Due to the foreign nature of the pathogen, the native population does not have the immunological defense to resist initial infection. The influenza is a quintessential example of such a disease, historically and in modern times.

In 1918 the first world war ended. Everyone was relieved that the conflict in Europe was over and the soldiers were coming home. Unfortunately, the world was marching into the next battle of 1918 [52]. This conflict was with the H1N1 Influenza virus also known as the Spanish Flu. This was the first pandemic that was extensively studied and had an unprecedented mortality rate that is higher than any influenza pandemic that has occurred since. It was calculated that over 2.5% of the cases resulted in fatality, while the typical influenza pandemic results in an approximate 0.1% fatality rate [31, 43, 52]. The graph below demonstrates the presence of the illness in the population [52].

Number of Cases compared to the
World Population in 1918
[per million]

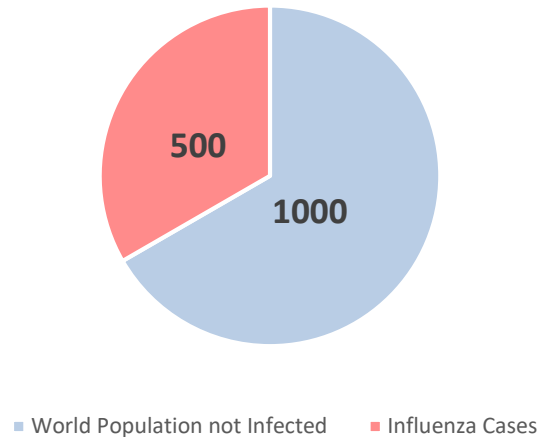


Figure 1 Number of Influenza Cases in the Population

The pandemic died down 18 months after the disease's initial outbreak. Modern containment efforts, aided by technological improvements in the communication systems, diagnostics, treatments, and sanitary refinements were particularly critical in containment efforts for this pandemic.

The 2009 Influenza broke out in the North America. It was a mutation of the same H1N1 virus, technically classified as the H1N1pdm09 virus or its more common name, Swine Flu. It was dangerous because the primary precaution, the influenza vaccine, was not extremely effective for this strain of the virus. This failure was not the result of lack of preparation but rather targeting the incorrect strain. The world was building their defenses against another, the H5N1 strain, or avian influenza, which was thought to be the immediate threat rather than Swine Flu [54]. The effective vaccine was not available until months after the vaccine's initial creation [21]. Throughout April and May 2009 various actions were taken by the CDC and WHO, including increased distribution of antivirals, distribution of the correct vaccine, a tracking log that healthcare professionals updated when diagnosing the disease to monitor it as it spread, public safety campaigns and social media awareness that aided the previously discussed disease

tracking. The pandemic eventually became under control once the actions were set in place [54]. These timely actions from the healthcare community helped limit the impact of the disease and increased the quality of care for all those affected. This can be seen when directly comparing these two pandemics side by side.

Error! Reference source not found. Figure 2 Prevalence of the Flu in 1918 and 2009 below outlines the differences between the 1918 influenza outbreak versus the 2009 influenza outbreak:

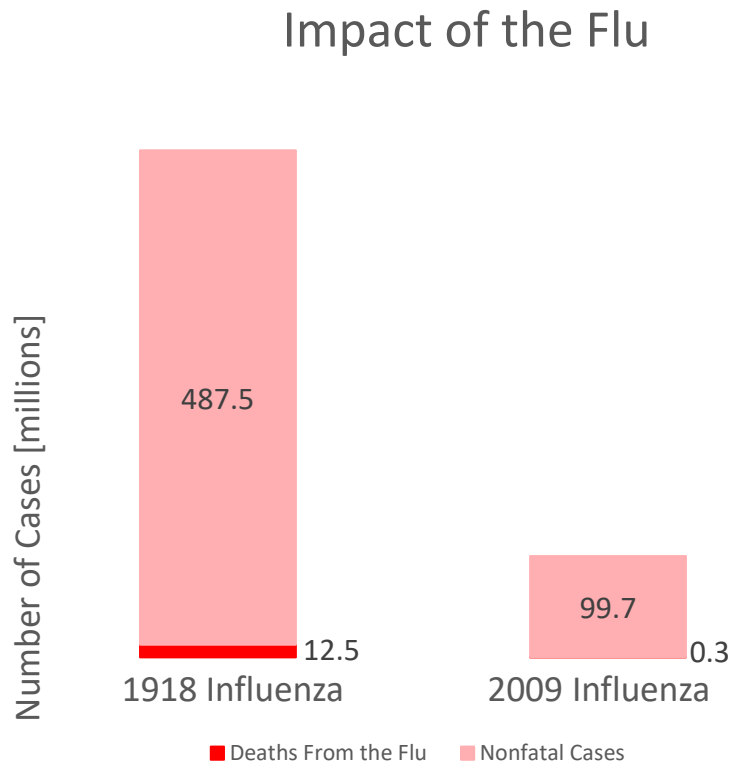


Figure 2 *Prevalence of the Flu in 1918 and 2009*

These pandemics demonstrate that the increase in technology and communication benefit disease containment. These results are even more impressive when considering the world population was approximately 4.5 times larger than in 1918 (world population was 6.8 billion in 2009 [62]). This can be viewed in Figure 3 World Population versus Flu Cases 1918 and 2009 below:

Population of the World verse Flu Cases

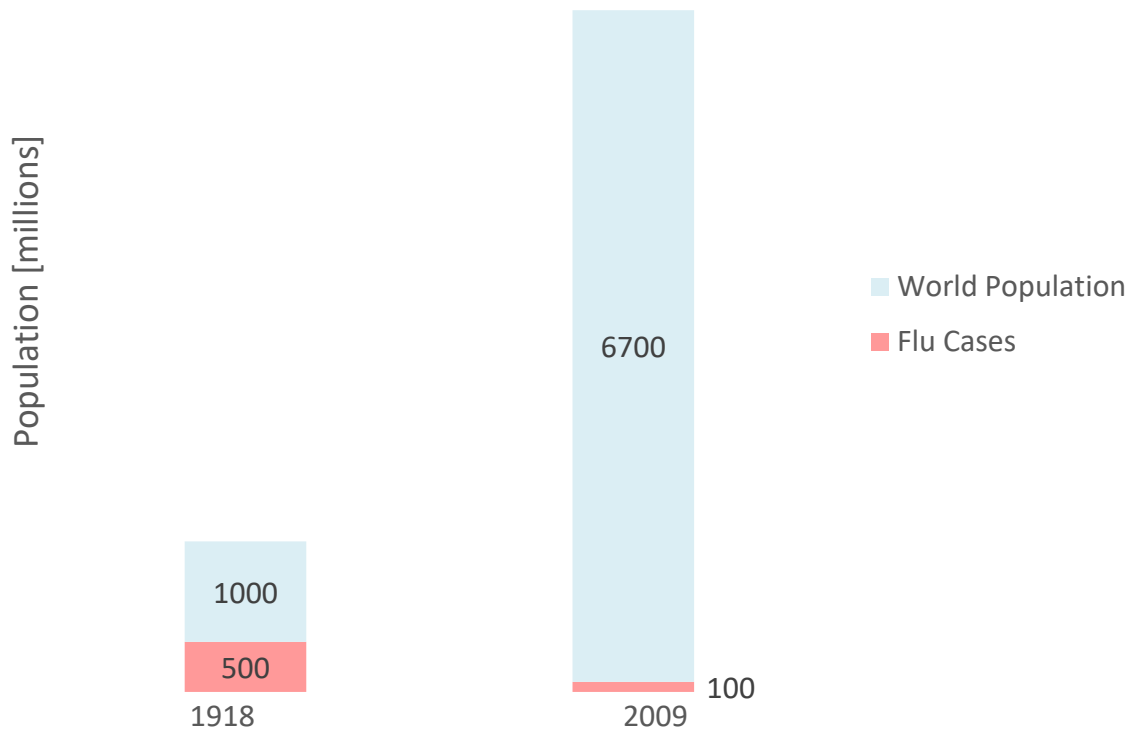


Figure 3 World Population versus Flu Cases 1918 and 2009

The 1918 Influenza pandemic demonstrates how truly terrifying and impactful a pandemic can be on a large portion of the population. The 2009 Influenza pandemic demonstrates the power in the current methods for treatment, however, it also demonstrates that the disease can spread rapidly if the world's current controls fails.

These measures have progressed in both speed identifying pathogens and creating treatments. They reduce the risk, however, the controls are not perfect. **Benjamin Franklin** said it best when he stated:

“An ounce of prevention is worth a pound of cure.”

This quote taken slightly out of context (Franklin was talking about fire prevention when he said it), however, it is extremely accurate for the modern diagnostic technology as well. In this thesis, multiple technologies are used and assembled into various experimental systems to fundamentally advance the current state of the art for noninvasive detection of illness using febrile status as the primary indicator of health. The decision for use this tool is due to fever being a very common symptom of illness [16].

1.2. Thermodynamics and Physiological Heat Transfer

Thermodynamics, more specifically physiological heat transfer, is the primary mechanism for fever manifestation. This section introduces a brief overview of thermodynamics and how the body uses this process to allow our bodies to run as efficiently as possible.

1.2.1. Thermodynamics as Applied to the Body

Heat is an output of any given system. It can be intentionally introduced into the system or be a byproduct of a desired function. Thermodynamics is the broad subject that characterizes intentional and non-intentional means of heat displacement. There are three laws that are the primary framework for this; the first, second and third laws of thermodynamics.

The first law of thermodynamics states energy cannot be created or destroyed in an isolated system. Due to the body being a semi-isolated system, in every physiological process, energy (primarily chemical energy fueled by nutrients) is translated into work and thermal energy. When considering the body, the internal body can be considered a closed system, even though there are some loss to the surface, while considering the external surface of the body it is an open system, with interaction with the ambient environment.

The second law of thermodynamics states entropy of any isolated system never decreases. Entropy is an abstract concept that states thermal energy is equal to the variation from one system to another. This means that entropy is the degree of thermal disorder in the system. The body is not perfect so this means that they are in a constant state of energy transfers from one physiological system to another.

The third law of thermodynamics states that as the temperature of a system approaches absolute zero the entropy of the system will approach a constant value. This applies primarily in the inverse; that is, as the body responds to diseases the entropy of the system increases and the temperature of the body rises.

As the various organs and systems of the body have optimal temperature ranges for their health and operation, heat is transferred to the surface of the body and released to the environment to limit hypothermic and hyperthermic conditions. Physiological Heat Transfer is studied in extreme detail to ensure high precision modelling to allow internal body temperature to be measured from surface temperature.

1.2.2. Physiological Heat Transfer

Adaption to adhere to the Laws of Thermodynamics was a necessary evolutionary advancement. The body optimized thermal detection to be capable of identifying the optimal temperatures for different organ function, and control localized temperatures to stay within the operational regions. This is the basic premise of thermoregulation inside the body.

Heat is a byproduct of organ function and it is produced by metabolic action. This metabolic heat is detected by thermoreceptors inside the body [10]. The physiological receptors can either be hot or cold receptors that activate at various frequencies, and deliver their information to the preoptic area and the anterior hypothalamus (also known as the body's thermostat). Both regions cannot be controlled by normal human cognition and as a result are categorized as part of the autonomic nervous system. The cold thermoreceptors begin to signal at 33°C (91.4°F), have their peak response at 28°C (82.4°F). They stop firing at 10°C(50°F), this is the point where humans feel numb. Warm thermoreceptors begin firing at 35°C (95°F) and are located at a slightly deeper level in the epidermal layer of the systems. System response time typically takes anywhere from 3 minutes to 30 minutes depending on the factors such as thermal conductivity, time exposed, insulative effect of the skin and adipose tissue etc. [10]. **This means that it can take almost a half hour for the physiological structure to acclimate and reach thermal stability!**

Once the thermoreceptor activates, the signal (known as an action potential) propagates to the preoptic area and the anterior hypothalamus. These regions of the brain send another signal out to impede metabolic generation of heat and provide mechanisms for heat loss in a controlled manner or increase the localized metabolic rate to increase heat generation in the area while increasing thermal isolation of that area [10, 56]. An example of this is the vasodilation that is experienced if an individual is exposed to a warm ambient temperature. The body's goal is to decrease the insulative effect of the skin and use the free energy resource (the warm ambient temperature). While vasoconstriction takes place if the individual is in a cold ambient temperature [56].

Another example of this is the heat byproduct produced by muscles when exercising. This is removed from the body through perspiration. Liquid perfuses through the pores of the skin, and with the excess internal heat changes phase to a gaseous state, decreasing the temperature of the body [10].

To simplify the study of thermoregulation of the human body it is separated into various subsections. These sections are: core temperature (including muscle interaction), and skin temperature (including the clothing layer). The clothing layer is the extra clothes a human wears to protect them from the ambient elements and the skin which has impact of the ambient temperature. This layer also serves as a protective insulative barrier for the second section, core temperature. The temperature of the organs that has to be kept at a safe level so that all systems function at maximum efficiency [20]. The external layer is an open system with many different complexities with the ambient interaction. This is also the primary reason for **reference devices** being semi-invasive. Measuring core temperature has to target a physiological orifice (i.e. the mouth, external auditory meatus, or anal cavity; all well recognized approximations of core temperature) or penetrate through the outer layer of epidermis (skin) to utilize the internal closed system and minimize the variables interacting in these reactions.

Blood serves as the primary mechanism for providing nutrients in the body (such as oxygen) and removing byproducts (including heat) from the body [20]. At rest, approximately all the blood circulates throughout the body over the duration of a minute (5-6 Liters per minute). Circulation

is accomplished by forces acting on the arteries from the smooth muscles surrounding them. Veins, with their lack of surrounding muscle, act as blood revivors. Since there is an exchange of nutrients via blood at the capillaries (area where veins and arteries meet) the arteries deliver the blood to the capillaries while veins take the blood from them, therefore, arteries are in a constant state of cooling while veins are in a constant state of heating up [20]. This is an integral factor when deciding a site to take a temperature reading from.

The core-to-skin transfer coefficient is a characterization of the site's ability to acclimate to core body temperature [20]. Due to the impact of blood flow and the dynamic nature of the blood vessels in thermoregulation, the core-to-skin transfer coefficient correlates strongly with proximity of blood vessels. There are many different factors to be considered when characterizing local temperature. Depending on the site, the impact of the factors varies and can be idiosyncratic to the individual. Relevant factors can be age, digestion status, ambient temperature acclimation status [20].

Febrile reactions can be characterized as an increase in core body temperature. Different diseases have different reactions and disrupt various processes. Fever can be a symptom of bacterial infections, certain viral infections (such as the flu) and inflammation [44].

Historically there have been various means for characterizing core body temperature, these will be discussed in the next section, section 1.3 Human Temperature Measurements.

1.3.Human Temperature Measurements

The human body is a very complex dynamic system that beautifully balances a wide variety of factors. Heat was one of the first metrics of the early physicians to characterize health. This section will outline the history of characterizing physiological thermal responses all the way up to the modern day.

1.3.1. Brief History

The inception of the measurement of human body temperature dates back to ancient times. "The earliest references to fever appear in Akkadian cuneiform inscription dating from the sixth

century BC and the context used is referring to magic" [10]. Hippocrates was next, and is credited as the first person to be recorded to explain fever is an excess of heat in an animal [10]. In these ancient times the heat omitted from an animal was recorded by the physician physically touching the infected site [10, 11]. This practice then progressed until the invention of a thermometer took place.

The first thermometer's purpose was to measure the heat of the room where it was held. It utilized thermodynamic responses of increasing and decreasing density of oil to cause the rise and fall in orbs held inside water [4, 11]. Daniel Fahrenheit took this model and used a more reactive substance, mercury, to then create a scale (in 1724) to quantify various physical phenomena such as the boiling and freezing point of water (212 and 32 degrees F) and human body temperature (initially 96 degrees F taken from the axillary site from his wife) [4]. Very soon after Fahrenheit's scale was established, Andres Celsius established a scale (in 1742) that set the freezing point of water to 0 and water's boiling point to 100 [7]. To this day, the United States still uses the Fahrenheit scale to measure temperature, while the common temperature scale used by most other nations is the Celsius scale.

At this point in history thermometers had just started to be used on humans. The next major advancement for this research came in 1868 by physician Professor Carl Wunderlich, and his treatise "Temperature in Diseases" [11, 57]. In this treatise, Prof Wunderlich outlined a set of rules defining the value of human body temperature measurement as follows:

1. *"The average normal temperature of the healthy human body in its interior or carefully covered situations on its surface varies, according to the plan of measurement from 98.6 to 99.5 °F (37 to 37.5°C)*
2. *A normal temperature does not necessarily indicate health but all those whose temperature exceeds or falls short of the normal range are unhealthy.*
3. *Alterations of temperature may be confined to special regions of the body which are the seat of diseased actions (local inflammation) while the general temperature remains normal.*
4. *Exceedingly low temperatures are very commonly met with in the following cases:*

- a. *In the remission of a remittent fever.*
 - b. *in consequence of loss of blood.*
 - c. *Sometimes in the death struggle. Abnormally low temperatures may seriously disturb the various functions of the body, and may render the continuance of life impossible.*
5. *Temperature can neither be feigned or falsified. It furnishes a certain proof of the reality of death, where this is otherwise uncertain” [11, 57]*

Prof. Wunderlich's was so accurate with these clinical rules, that they are still in practice today; normal temperature is still denoted as a range, lack of fever does not mean perfect health, localized hot areas are still used to denote localized infection, temperature still cannot be falsified easily and repeatedly.

Upon the invention of the microcontroller, digital stick thermometers began to appear on the market. Initially not as publicly welcomed as the glass thermometers, the digital stick thermometers began to capitalize from their digital nature and utilizing predictive measurements. This cut down the time needed to take a measurement from 3 to 15 minutes (depending on the site measured) to 1 minute (with a digital stick thermometer) now as fast as 2 seconds [37].

Next came the tympanic infrared designs, which were an improvement from the predictive digital stick thermometers, but still semi-invasive to the patient due to their need to be inserted in the external auditory meatus of the patient. This allows for characterizing of the heat emitted from the tympanic membrane. It is believed that this is such a stable and reliable location due to its proximity to the internal thermostat (anterior hypothalamus).

The latest advancement in home use and clinical thermometry is noninvasive infrared measurements. These measurements have susceptibilities of increased impact of the factors discussed in section 1.2.2, however, it is completely non-invasive to the user, making it ideal to measure a sleeping child or to maximally limit complications from non-biocompatible materials or transmission of disease on the surfaces of the device. The basic premise of the radiation

based measurements will be discussed in section 1.4 Radiation Based Measurements and the advantages and disadvantages of Tympanic measurements verse Forehead measurements will be discussed in section 1.5 Infrared Measurements in Practice.

1.4.Radiation Based Measurements

Radiation based measurements of facial temperatures are the primary focus of this research. This is the monitoring of energy that is radiated from the face at distinct wavelengths that are indicative of the energy created by a human body.

1.4.1. Wavelength Basics

All visible light is a form of a quantum particle that is resonating at a specific wavelength. Human vision is limited to a small subset of this spectrum (Figure 4 Electromagnetic Spectrum [48] below)

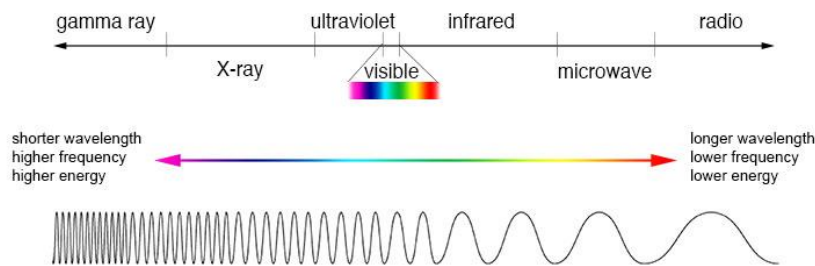


Figure 4 Electromagnetic Spectrum [48]

An important wavelength that cannot be seen by humans is infrared radiation. Infrared radiation (IR) is the primary subset of the electromagnetic spectrum that heat resonances at. The peak wavelength for the range of the typical temperatures emitted by human skin is approximately $9\mu\text{m}$ - $12\mu\text{m}$, however, measurement devices from $2\mu\text{m}$ to $15\mu\text{m}$ have been successfully used in recording thermal radiation of physiological structures [39].

1.4.2. Infrared Radiation Sources

The first law of thermodynamics states that energy cannot be created or destroyed, therefore, heat is a byproduct of metabolic reaction(s) (as previously discussed in section 1.2.1). Since the

heat is not needed it must be removed before it has detrimental effects on the health of the individual. The devices are monitoring the transfer of heat from the sources.

1.4.3. Means of Detection

The primary sensor used for the characterization of infrared radiation are thermopiles.

Thermopiles detect the radiation from the difference (Δ) of charges of a given material when the material is at rest and when radiation that has infrared wavelengths act upon it. These materials change electrical dipole moments from the directions they are aimed [55]. To further simplify this means of detection, if the thermopile transducer is aimed at a target that is emitting heat, the waves propagate to the sensor causing a change in the material's charge. The primary characterization of the changes is the charge (voltage) that is sent from the sensor.

Since all forms of light are electromagnetic radiation, manufacturers of these devices typically place a material in front of the transducer that allow the intended infrared wavelengths to pass through and reject less desirable wavelengths. This material acts as a bandpass pass filter to minimize the effect of different spectrums, (such as stray light) from being detected and maximize the infrared signal. A graph of the bandpass wavelengths of a HgCdTe alloy (a material with the correct filtering properties) can be seen below:

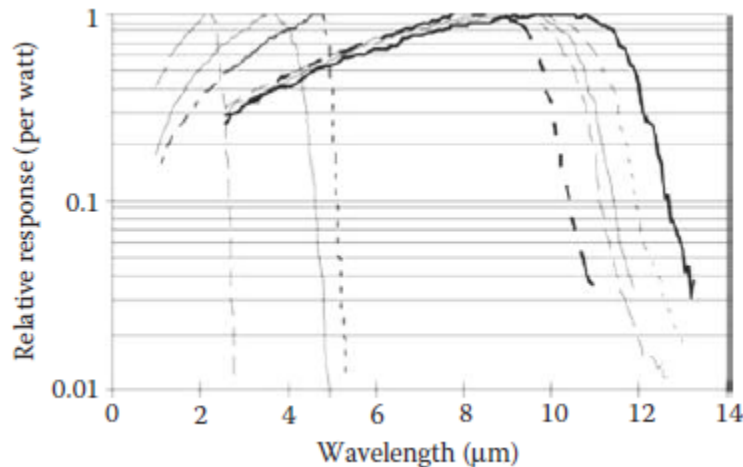


Figure 5 Spectral responses for different HgCdTe alloy detectors [35]

This phenomenon is analogous to rod and cone receptors in the human eye. The rod and cone receptors allow certain wavelengths of color to be seen, typically characterized by red, green and blue. One of these filters only allow for the detection of a single color. The primary difference between the rod and cones of the eye and this material is the wavelengths they allows to penetrate (infrared verse visible light). The radiation permitted by the filters are then used as optical input based upon the number of sensors.

These sensors then can be used as a single discrete measurement or be combined into focal point arrays of thermopile sensors. Section 1.5 Infrared Measurements in will outline its current use in medicine and as a screening device.

1.5. Infrared Measurements in Practice

There are a variety of different clinical diagnostics and applications for which IR sensors have been used due to their noninvasive nature. For the sake of this research the primary focus is going to be Temperature measurements.

1.5.1. Core Body Temperature with Thermopile Sensors

There are a variety of different products that utilize various physiological locations to characterize core body temperature.

One of the most precise forms of measurement of core body temperature is tympanic measurements. Tympanic measurements are comprised of a thermopile sensor some distance inside the external auditory meatus (ear canal), recording the infrared emission from the tympanic membrane or the immediately surrounding tissues [23]. When the device is in contact with the human body a heating or cooling element can be added to reduce the amount of thermal shock (incorrect measurement due to rapid heat transfer from the surrounding tissue to the device), increasing the accuracy of the device [22, 24]. Currently, the industry leader of this category are the Braun Thermoscan devices.



Figure 6 Welch Allyn Pro6000 Tympanic Thermometer

Another way that IR technology is used is via forehead measurements. This gives the operators (fortunately or unfortunately depending on the intended use of the design) the option for a dynamic measurement of a single structure or a single static measurement. Some devices target several locations across a single blood vessel structure [53]. The real benefit of this approach is the minimization of user error due to the incorrect site being targeted. The processing is typically a function that then takes the raw temperature reading of the site and the ambient temperature to output a reading equivalent to core temperature based off the characterization of the area seen. An example of this device is the one that swipes across the forehead that can be seen in Figure 7 Exergen Temporal Thermometer Use [53]:

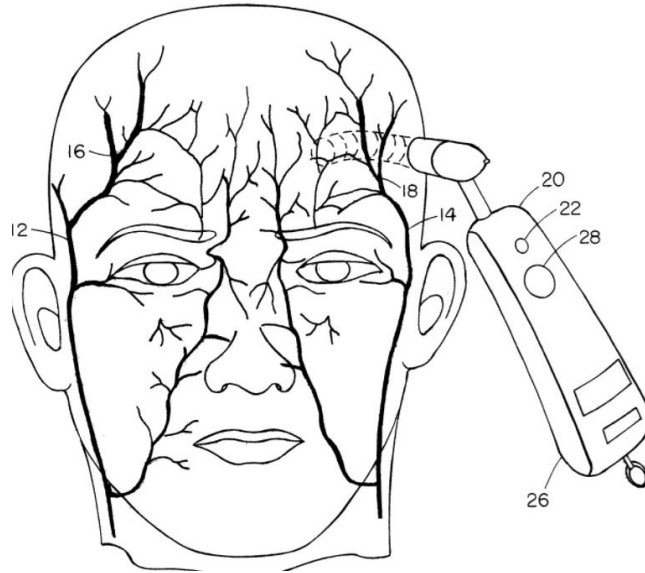


Figure 7 Exergen Temporal Thermometer Use [53]

Another means of detecting forehead temperature is by enhancing the signal using a parabolic mirror to increase the amount of IR radiation directed to the sensor. This allows the area subject to measurement to be kept at a near uniform size independent of distance. This can be supplemented with the addition of a distance sensor to the device to apply a compensation for the loss of the system overall [58, 60]. One device that utilizes both of these approaches is the Braun NTF3000, seen in Figure 8 Braun NTF3000 [36]:



Figure 8 Braun NTF3000 [36]

These are the two primary means of measurement for forehead devices with a single pixel measurement, however, they do have their drawbacks. They are dependent on the device and subject to undergo thermal acclimation with ambient conditions, they are dependent on single points that are only targeting a single physiological structure (the scanner takes the maximum or average value of a single structure, therefore, targeting a single area) and they are very susceptible to the skill of the operator with very little 'forgiveness' for incorrect use.

Thermal arrays have the potential of overcoming these hurdles.

1.5.2. Core Body Temperature with Arrays

The natural evolution of technology is moving toward fever characterization with IR Focal Point Arrays (IRFPAs). These are arrays of thermopile transducers that allow the creation of digital images from the IR signal. The downside of this type of sensor is the curse of dimensionality, meaning there are too many data inputs. Each input applies a new dimension to the data, and potentially a different feature that is present. For this reason, more sophisticated feature extraction techniques are needed to:

1. Extract the features and
2. Use the features to render the value

Ring, Jung, Kalicki, Zuber and Vardasca found in 2012 that 9 pixels located in each corner of the eye to provide an indication of fever was sufficient for targeting this specific feature [39]. Other methods typically entail obtaining the axilla) in combination with another physiological location such as the inner canthus of the eye to create a stable consistent measurement [40].

More reassurance that thermal imaging is the correct tool for fever screening came during the Severe Acute Respiratory Syndrome (SARS) epidemic. Chiu et. al used a thermal imager to characterize fever and help the ill get the help they needed using the maximum intensity value in a near static ambient environment [9]. Necessary for Chiu's method was precise ambient control, precise positioning of the subject in the frame, and thorough acclimation of the subject to the ambient temperature.

This technique is known as a **classification output**, which outputs a discrete value indicating if the input is febrile or afebrile. The previously discussed temperature readings that people conventionally think of are known as a **Regression model**, outputting a value in the Celsius or Fahrenheit scale. The **Classification model** give some distinct improvements:

1. It allows for more flexibility in the scale of the data inputs.
2. It relays less information. When temperature readings are taken to aid containment efforts aimed at a pandemic or similar disease, what is desired is a confirmation of illness or health rather than an estimation of severity of illness from a continuous temperature reading.

The **Classification model** is typically the model chosen for screening techniques due to these advantages. This is ideal for identifying and assessing large volumes of people in a short amount of time to minimize the risk of diseases spreading to the general public in high volume locations, such as airports, which can introduce new diseases to new populations and spread very quickly.

1.5.3. IRFPAs as Screening Devices

There are various studies emerging that incorporate IRFPAs in combination with algorithms of varying complexity to create a system that can characterize the health status of an individual with fewer constraints to allow easier and more accurate implementation.

In 2017 the International Organization for Standardization (ISO) came up with standards for fever monitoring to combat the threat of pandemics around the world. In the standard titled "IEC 80601-2-59:2017 Particular requirements for the basic safety and essential performance of screening thermographs for human febrile temperature screening" there are a variety of general requirements outlined. These include:

- Removal of obstructions (such as hair and glasses) [clause 201.7.9.1]
- Recommendations for ambient temperature between 18°C and 24°C and a relative humidity of below 50% (and the subject should not be sweating) [clause 201.7.9.3.1]
- Near real time algorithms [clause 201.12.2.102]

- Minimal target plane of 320x240 [clause 201.12.2.103]
- Face encompass 56% of the minimum display of the target plane [clause 201.12.2.103]

There are a variety of other standards and clauses that may apply, however, these are the primary ones that will be addressed.

In 2008 Bitar, Goubar and Desenclos conducted a review of the effectiveness of different sites in the prediction of core body temperature. The following table is what was found:

First Author, year	Sample Size	Target area	Temperature threshold [°C]	Fever Prevalence [%]	Sensitivity [%]	Specificity [%]
Ng E 2004 [33]	310	Forehead	37.7	16.9	89.6	94.3
	310	Inner eye corner	37.7	16.9	85.4	95
Liu 2004 [29]	500	Forehead	37.5	unknown	17.3	98.2
	500	Auricular meatus	37.5	unknown	82.7	98.7
Chan 2004 [8]	188	Forehead	38	14.3	4	99
	-	Forehead	37.5	N/A	15	98
	116	Auricular meatus	38	20.7	67	96
Ng 2005 [32]	500	Forehead	37.5	12.3	89.4	75.4
Chiu 2005 [9]	993	Forehead	37.5	1.2	75	99.6
	72.327	Forehead	37.5	-	-	-
Hausfater 2008 [15]	2.026	Forehead	38	3	82	77

Table 1 Bitar, Goubar and Desenclos Research Comparison [3]

While having large sample sizes, these studies have a low prevalence of fever. This makes it difficult for the algorithm to learn the correct information from these sites that have a significant level of variation due to the physiological variables (previously discussed in Section 1.2

Thermodynamics and Physiological Heat Transfer).

In 2014, Sun et. al explored the feasibility of detecting fever with a smaller array (48x47 pixel array). In their study, the system was placed 30cm from the subject, in an ambient range of 72°F-74.8°F (22.2°C-23.8°C) and a relative humidity of 36-40%. The binary result of febrile status was compared to the readings from an axillary reference device, with reference febrile status being defined as an axillary temperature value greater than 37.5°C (99.5°F) [50]. 36 patients were found to be febrile. The algorithm used was a simple threshold; if the max pixel was above 97.7°F (36.5°C) then the patient was said to have a fever. The sensitivity was 80.5% and specificity was found to be 93.3% [50]. This challenge to the standard demonstrated that the high resolution array is not required for fever screening

One of the most successful monitoring studies that has been conducted is by Professor Ng of Nanyang Technological University. He created an algorithm that uses the K-means clustering algorithm to create a radial basis to find locations that are then inputted into a Fully Connected Artificial Neural Net for feature classification. This complex algorithm was able to output with 96% accuracy (claiming a sensitivity of 100% and specificity of 94%)[34]. To obtain these results his algorithm necessitates retrieving data from the same patient facial orientation, with a fixed distance, in a stable ambient (temperature of $25 \pm 2^\circ\text{C}$ with a relative humidity of approximately 60%), with a FLIR IR ThermoCAM S60 system that has 320x240 pixels resolution[34]. This algorithm, despite the constraints required in use, demonstrates the potential for classification algorithms for fever screening.

The current state of thermal imaging for biomedical monitoring has the foundation built and ready for further growth. This foundation entails devices that output core body temperature based upon a given area of the face. This approach is susceptible to user bias and systematic

variation. This research aims to reduce the unpredictability due to various forms of variation with flexible algorithmic design, utilizing Machine Learning.

1.6. Brief background of Machine Learning

Artificial Neural Networks are biologically inspired algorithms that output classification or regression values of a given input data. These networks only output one of these two possible outputs based upon their specific learning algorithm (often square error for regression values or cross entropy for classification). This technique is so powerful it is now the building block of modern Artificial Intelligence.

The first publication that made this type of learning viable was published in 1986. It was Rumelhart, Hinton and Williams research, titled, "Learning representations by back-propagating errors." In this paper the team outlined the guidelines on how to create an artificial intelligence network; first run the data through a network of matrices and **tensors** (multidimensional matrices) and allow the results to converge at the end. The error is found at the output (using nodes with initial weights and biases named neurons) and traced back to the front of the network, modifying the neurons that contained weights and biases that caused the error. This is then repeated until the optimized weights and bias are learned from the input data [45]. This is known as the backpropagation algorithm.

The next big breakthrough occurred in 1998 by LeCun, Bottou, Bengio and Haffner. They discovered a technique involving the process of convolution in an image, or multiplying a given area of pixels by a filter (with the same area) and outputting the sum value in a new matrix. The filter is then slid across the image (or image across the filter). This type of network can then train the filters with the backpropagation algorithm, similar to how the nodes of the neural network were optimized before [28]. This made the classification of images simpler, and could be used on things such as number identification.

The next major advancement occurred in to 2012 with the inception of AlexNet in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). This Convolutional Neural was a major breakthrough in the current state of the art, outputting a misclassification rate of 15.4%, beating

the second place by over 10% (second had an misclassification rate of 26.2%). This started the deep learning boom for object detection

For the next subsequent years this type of feedforward CNN won the challenge, until 2014 where VGGNet lost to GoogLeNet [51, 52]. This began a new turning point in deep learning from conventional linear feedforward architectures, to creating networks that use different layers to obtain higher classification accuracies (when sufficient data is available). In the effort to improve efficiency in development of new applications, it was further found that the use of pretrained deep networks to create new classification and regression based systems decreased the data dependency of the technique.

1.7. Deep Learning Functionality

1.7.1. Neuron

The basic building block for deep learning is the neuron. The conceptualization of a neuron is biologically inspired nodal points that are interconnected [17]. These nodal points all have specific weights and biases attributed to them. These values are applied to data as it is propagated through the neurons. If the resulting values are large enough to exceed a learned threshold, the signal continues.

The resulting propagation of the signal from the application of the weights and biases of the neurons is determined by an activation function. The most common activation function used for the hidden layers of an architecture (middle layers that are neither the input or output) is a rectified linear neuron (or ReLU) [18]. The equation for this activation function can be seen in Equation 1 ReLU Activation Function below:

Equation 1 ReLU Activation Function

$$\text{output of neuron} = \text{bias} + \sum \text{signal}_i * \text{weight}_i$$

The final activation function, used to classify the data, is known as the Softmax function.

This equation has a sigmoid nature that allows for predictions on a binary scale for a given class and is artfully designed to allow multiple different classifications in addition to a simple binary output, allowing the network to scale for multiple prediction types. The equation for the Softmax activation function can be seen below:

Equation 2 Softmax Activation Function

$$Classification_{single\ input} = \frac{e^{output\ for\ an\ input}}{\sum_{number\ of\ classes} e^{output\ for\ a\ class}}$$

These are the only two types of activation functions used in this research.

In execution, neurons are assembled in groups known as fully connected layers to align with the size and format of the data that is being analyzed (for example, the very first fully connected layer would have the same number of neurons as pixels in the input image, adjusted for any resolution reduction or cropping performed on the input image) and the weights and biases for any given fully connected layer are updated during the learning phase to optimize the accuracy and performance of the prediction. This learning is accomplished, in large part, with the use of backpropagation.

1.7.2. Feedforward Networks and Backpropagation

Feedforward networks are deep learning architectures where each layer of neurons is layered to allow the architecture to be executed as data moves forward through the architecture while maintaining backward traceability (computers can easily trace the error from the output to the input, while it would be extremely difficult for a human). Rumelhart, Hinton and Williams's research with the application made the training of these architectures possible. This advancement allows the network to target the neurons that make the incorrect classification and modify the weights, resulting in less error.

Backpropagation is running an input(s) forward through the network, calculating the error at the output, and slowly moving back to the input layer, computing the derivative of the error at each neuron, update the weights and biases of these point as the algorithm progresses to the input layer [18]. The power in this technique is the ability to learn the computed error for each neuron (nodal point) in the architecture when it is allowed to be trained. To compute this error a powerful equation is needed. This equation is known as the loss function of the network. For this research, binary cross entropy is used, and the equation for this can be seen below:

Equation 3 Binary Cross Entropy

$$loss = -(weight_{fever} * output * \log(probability) + weight_{afebrile} * (1 - output) * (\log(1 - probability)))$$

The weights of the neurons are optimized by minimizing the absolute value of this loss equation, with the goal being a loss of 0. This is how artificial intelligence is applied to deep learning.

There are also optimization functions that allow for increased speed for learning by modulating the size of the changes in the weights of the neurons to allow for larger changes when the loss is greater and smaller changes as you approach zero, also referred to as modulating the speed of gradient descent. These equations have variables (hyperparameters) that allow fine tuning for the application. In this research the adaptive moment estimation (Adam) optimizer was used [24]. Combining these tools in multiple layers is known as deep learning.

1.7.3. Regulation

In the design of a deep learning architecture, regulation of the network is limiting the network's ability to learn overly specific features from the specific inputs so as to force it to learn general features instead. It is an important point to note that the goal at the end of the training is a generalized design that is capable of theoretically characterizing all use cases, or a network that best fits the application. Overfitting, which happens with a poorly regulated network, biases the design to only those use cases that the network was trained from. An example of overfit can be seen in Figure 9 Visual Demonstration of Overfit.

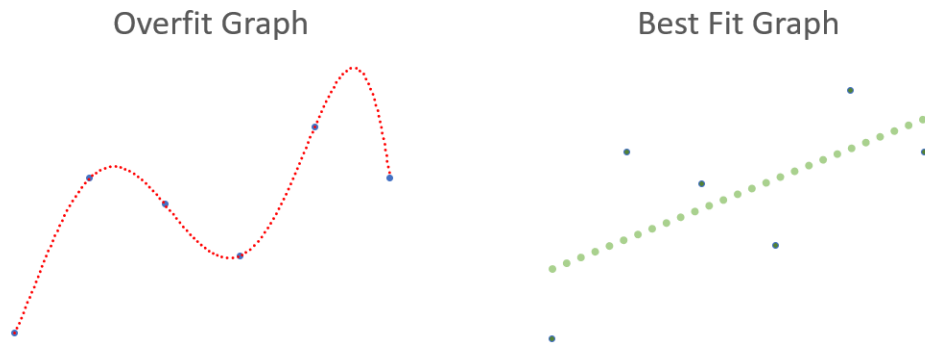


Figure 9 Visual Demonstration of Overfit

The importance in network regulation cannot be emphasized enough as this investigation was with a small dataset (the dataset will be discussed in section 2.1). There are design choices that Deep Learning Architects use to minimize overfit such as: separating the test and training set, addition of dropout layers and an L2 weight layer.

The test and training set must be separated to demonstrate if there is an overfit issue; by training the algorithm with one set of data (the training set) and validating it against a separate set of data (the test set) you ensure that the algorithm is trained to identify general characteristics rather than overly specific idiosyncrasies. The most common way to separate a dataset is by randomly dividing the dataset into a various number of folds, training on all but one fold and evaluating the results against the fold not trained upon. This is repeated for as many times as there are folds to ensure that each fold acts as the test set one time for a network trained on the other folds. This is known as k-Fold Cross Validation [41]. The k denotes the number of folds the data is separated into. This not only allows for the ability to visualize overfit, but a more accurate means of evaluating the validity of the network (if an additional clinical trial is not a viable option) as it ensures that the final performance of the network is a reflection of the actual network design rather than a random lucky result based on the specific combination of data used to train and test. For this research a 10 Fold cross validation was used. 10 was chosen because it allowed the maximum samples being allowed for training while a large enough sample size for validation.

Dropout layers can also be added if overfit is an issue for the architecture. These layers are placed before the learning layers (fully connected layers made up of neurons) and apply a probability of any given connection being turned off in training. This is effective because the learning algorithm will decrease weight to a given node when it is not used in a given backpropagation iteration [49]. If the connection for a given node is not activated during a round of training, the weight will not be updated or be used in the accuracy of the training of the current input. The back propagation algorithm will increase the dependency of the weights around the connection, decreasing the importance of the connection that is currently turned off. The probability of the given input being turned off was found to be 0.7 (empirically found to be the optimal value). This technique was supplemented with L2 regularization.

L2 regularization is applied a square difference from the current value and the estimate value (from the loss function, cross entropy) that is then multiplied by a scalar value. This technique will apply the shortest path to minimize the loss value. For this algorithm $3e-5$ was used for the scalar, which is less than the value cited by the original VGG16 network ($5e-4$).

1.7.4. Convolutional Neural Networks

LeCun, Bottou, Bengio and Haffner merged a developed technique known as convolution, or the merging of two signals, in conjunction with the concept of neural networks to create convolutional neural networks [28]. Small Filters (or kernels when applied to image processing) are dragged across the image, containing selected weights. These weights are designed to identify select features in the base image. The backpropagation algorithm is then modified to compute the error of all the weights in the filter and sum them together to create a feature extractor[12].

This technique is then supplemented with pooling layers. These layers reduce the dimension of the images (i.e. go from a 14×14 image to a 7×7 image). This reduces the dimensions and decreases training time at the expense of losing resolution of location and details of the data[12]. The output of these convolutional layers can be fed into multiple fully connected neurons to be trained, or by utilizing transfer learning, fed into another algorithm.

1.7.5. Transfer Learning – Pretrained Networks

Transfer Learning is a technique that allows for a reduction of the data by utilizing a network that has already been trained in one domain and applying it to the domain of the network that is being designed [25,58]. While using this technique it is standard practice to not train the layers that have been trained in the different domain [58]. This allows for the feature extractor to have defined filters that are capable of identifying select features of the image. When convolutional neural networks are applied in this research, they are left untrainable due to the data limitations of the study.

2.Data Gathered and Equipment

All materials outlined in this section were the property of Helen of Troy or the author.

2.1.Clinical Data

The data gathered for this research was predominantly obtained at clinical sites with formal methodology. A subset of the data was gathered in laboratory sites using the same methodology. Thermography data, using clinical screenings, was obtained by taking thermal images of the patients and evaluating these images against a reference temperature reading from a reference device to determine febrile status. These continuous values that were output by the device were then made in to discrete classifications, afebrile (the patient does not have a fever) or febrile.

A FLIR T660 Thermal Imaging System was used to obtain all gathered images for this research. For the specifications of this system refer to Table 2 FLIR T660 Specifications. The reference readings were gathered using a Welch Allyn Suretemp 690 Plus thermometer. Patients with a reference temperature greater than or equal to 99.5°F (37.5°C) were considered to be febrile

There were two different sites used, the primary site being Hospital del Niño Jesús in Tucuman, Argentina, and the secondary site was Helen of Troy's Healthcare Laboratory in Marlborough Massachusetts. The ambient temperature at the data gathering sites had varying values. The Helen of Troy site had an ambient of $62.6 \pm 3.6^\circ\text{F}$ ($17 \pm 2^\circ\text{C}$), and Hospital del Niño Jesús had an ambient temperature range of $80.6 \pm 3.6^\circ\text{F}$ ($27 \pm 2^\circ\text{C}$). The Hospital del Niño Jesús recorded image primarily of children (approximately 98% or 123 images). 77% of the dataset was from the Argentina site, which means 75% of the total samples validated were children (120 images of children). Helen of Troy subjects were all adults older than 22 years old (resulting in 37 adults or 23% of the total dataset).

The data was collected from 11 febrile patients, 104 afebrile patients. The data was taken with subjects at different thermal acclimation times, angles, distances, and emotional states (the dataset has 3 young patients crying). In total there were 810 images taken with 141 images

taken on febrile subjects and 669 images taken on afebrile images. Following collection, the data was then filtered to remove any images that had the medical professionals obstructing the physiologic structures of the face. The delimiters for each study will be stated in their applicable chapter.

2.2. Thermal Imager

Specification	Data
Resolution	640x480
Thermal sensitivity	<20mK @86 °F
Field of View	25 degrees x 19 degrees
Image Frequency	30Hz
Object temperature range	-40 to 302 °F
Accuracy	1% of reading for limited temperature range
Operating temperature range	5 to 122 °F

Table 2 FLIR T660 Specifications

2.3. Computer Specifications

Specification	Data
Manufacturer	Gigabyte Technology Co., Ltd.
Primary OS	Windows 10
Secondary OS	Ubuntu 16.04
Processor	Intel i7-6700K
Processor Speed	4.00GHz

Cores	4
Logic Processors	8

Table 3 CPU Specifications

2.4.GPU

In order to enhance computational efficacy a GPU was used. The specifications of the GPU can be seen below:

Specification	Data
Manufacturer	NVIDIA
Type	GeForce GTX 980
Memory Clock	7Gbps
Base Clock	1126MHz
Boost Clock	1216MHz
Memory Config	4 GB
Memory Bandwidth	224 GB/sec

Table 4 GPU Specifications

3. Research Overview

This research has taken shape in 3 different experiments with the data outlined in Section 2

Data Gathered and **Equipment**. Each experiment had different primary and secondary objectives, but aimed to tackle the same base hypothesis.

Hypothesis

By utilizing modern machine learning techniques, a correct classification of the febrile status of the patient is possible independent of thermal acclimation time, emotional status, age, and ambient temperature range.

This proof of concept was evaluated with a combination consisting primarily of the F1 score but also including simplistic accuracy and the sensitivity and specificity of the algorithms. The F1 score was set as the primary metric for evaluation with a value to demonstrate proof of concept at 0.8. The F1 score was selected due to it taking into consideration precision and recall [37]. Precision is the amount of times the algorithm was correct at identifying fever in relation to the amount of times it claimed fever was present. Recall is the amount of times the algorithm was correct at identifying fever in relation to the amount of times fever was present. This equation can be seen below:

Equation 4 F1 Score

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} = 2 * \frac{\frac{TP}{(TP + FP)} * \frac{TP}{(TP + FN)}}{\frac{TP}{(TP + FP)} + \frac{TP}{(TP + FN)}} = \frac{2TP}{(TP + FN + FP)}$$

With the expansion of the equation it can be seen that excess weight is given to outputs of positive predictions with minimal weight given to outputs of negative predictions [37]. The alternative metrics were considered to counter the risk of insufficient weight on false positive readings; however, in a device such as this, which is targeted to minimize the impact of pandemic scenarios, the risk of false negatives outweighs the risk of spending resources on a further investigation of individuals falsely identified as positive. At a proof of concept level, the F1 score gives the best balance of accuracy versus prevalence and risk. At a device level, further investigation would be needed to validate an ideal success metric.

This research is aimed at proving that the combination of these approaches is sufficient, at the proof of concept level, to address the noise factors in the data and not at a demonstration of a final algorithm. Each of these noise factors (ambient acclimation status, facial orientation, emotional state, febrile status, etc.) represent use cases that would occur for a device in the field with trained operators. The ambition of a final screening algorithm is to be common for the entire population. Validation of a population based method requires a substantial sample size that is not present in this research.

Initially the goal was to identify the most meaningful locations on the human face and then use the identified area or areas to generate a decision on the patient's febrile status (similar to the studies taking into consideration a single site). As the research progressed, the team abandoned this conventional approach in favor of classification of febrile status via inputting the thermal data into a neural network to allow the network to dictate the correct classification.

It was evident from the beginning that the gathering of the data was an issue. The data was gathered in without a common thermal acclimation time, common emotional status and in wide ambient temperature range , making the approach robust enough for real world applications. The uncontrolled use cases manifested itself as noise in the data.

The reason for overcoming these use cases is rooted in new technology development ideology. It is believed that the creation of an algorithm sufficiently robust to correct for errors in the supplied data is superior than simple identification and avoidance of the sources of error. The necessary evil of permitting noisy data allows for a more useful real world algorithm that is robust to correct for error. The different experiments conducted were as follows:

1. Localized Area Investigation (Chapter 4)

The primary objective of this investigation was to identify and extract temperature data from the features on the face to correlate linear relationships between these sites with the noise present. Instead of relying on a supervised segmentation algorithm to output the areas of interest, each image was masked by key facial features outlined in the publications depicted on Table 1 Bitar,

Goubar and Desenclos Research Comparison [3]. Descriptive statistics were extracted and inputted into a given equation and optimized for each site using a least squared error algorithm.

This investigation was ultimately unsuccessful due to it attempting to quantify linear relationships. The relationships are not linear in nature (at least in the presence of noise). The hypothesis was found to be incorrect due to the nonlinearity of the inputs, however, it directed the research to the exploration of nonlinear methods of evaluation. This raised the question, "are convolutional neural network flexible enough to apply a correct classification of fever?"

2. Binary Classification with Pretrained VGG16 (Chapter 5)

The primary objective was to create a fully supervised algorithm that would be capable of meaningful classification of the febrile status. The reason for the pretrained model is to combat the data issue (lack of data in combination of excess noise).

The manner that the data was randomized was integral to the reliability of the validation set. If the data was randomized by the **augmented data**, then the network would learn the features of the image resulting in outputs that are not a true representation of the performance of the network. For this reason, the inputs were randomized by the individual image rather than the augmented input. The imbalanced dataset was also a design consideration that was accounted for in the final design of the network. Overall, this experiment was successful, and the results were able to accomplish the goal. The final experiment's goal was now to improve upon these results.

3. Feature Extractor – PCA – SVM – Vote Algorithm (Chapter 6)

The primary objective for this experiment was beat the benchmarked value that was found in Chapter 6. To accomplish this the team elected to blend the benefits of the pretrained VGG16 Network, using the pretrained convolutional layers as a feature extractor, then reducing the dimensions of the data using principle components analysis (PCA) that then is inputted into a support vector machine (SVM).

This method is a well-studied approach, but new for this fever screening application [6]. It requires less data. If the VGG16 Feature Extractor can identify the important nonlinear features, the PCA should be able to identify trends and allow the SVM to make the classification.

This was the final experiment and for this data was found to have a better average result than the VGG16 network due to these qualities.

Disclaimer

This research utilized the Fahrenheit Temperature scale for the creation of the algorithms outlined in the subsequent chapters due to the following reasons:

- Displaying the temperatures in Fahrenheit scale allows for a slight increase in resolution for the data and
- This research was conducted by an American team.

From this point forward, all temperature data and results are displayed on a Fahrenheit scale unless stated otherwise.

4. Localized Area Investigation

The ambition of this experiment was to identify linear trends in the specific sites of the thermal images that can be distinguished in digital images. The first step in this approach was to manually mask and analyze areas of interest to evaluate if excess variation was present in the segmented structure, and evaluate if this structure outputted a linear (or quadratic) response. This approach mirrors the preceding work cited in Table 1. This is shown in Figure 10 Focus of Localized Area Investigation.

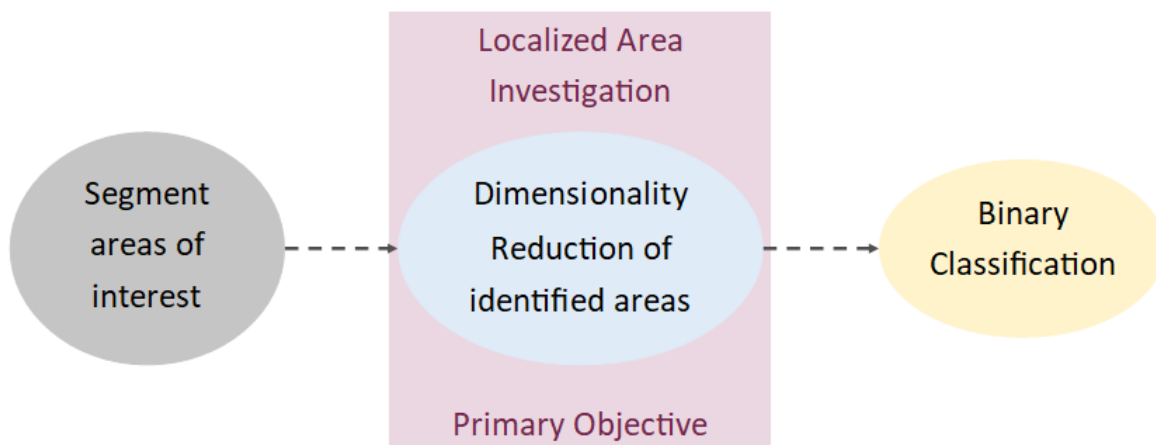


Figure 10 Focus of Localized Area Investigation

The optimal output from this investigation would be a single superior site giving a stable and linear output in comparison to the other sites. From this investigation, as will be shown, it became evident that there is no single feature that was superior.

4.1. Purpose

Many publications and devices outlined in Chapter 1 are EXTREMELY dependent on the concept of a single site's superiority. If there was a high correlation with these sites and the reference temperature, some function of the distribution should be able to be inputted to a transfer function and output an accurate reading. If this site could be identified, it would be extremely powerful and maintain the segmentation approach as a viable research route. This investigation was meant to provide some insight on the sites chosen and provide input on which

physiological feature can provide input to a classification network to determine febrile status of the individual.

The first step was to identify the sites to evaluate. It was decided to extract the left and right canthus (corners of the eyes), left and right temple and the center of the forehead so as to target the supraorbital blood vessels based on the current publications and thermometers on the market. If the structures were not located in the images, then there was no data recorded for the individual site and no training or testing input was provided for the given site. For each of the patients, at least one of the canthus, temples and the center of the forehead were able to be masked and extracted. For the forehead and temples, a diameter of approximately 1 inch (2.5cm) was used as the area evaluated. This can only be approximated due to the varying distances in the dataset.

55 images were used for this analysis with 18 being from febrile subjects. The means of selection of subjects were random, and the primary output was a continuous core temperature regression value. The purpose of this experiment was to be able to identify on a subsample selection of the total population if there was any standout feature. Basic linear correlation was not a viable option due to the bias of the sites vs core body temperature having a relationship with ambient temperature. Multiple equations were evaluated and it was found that a quadratic relationship between site temperature and reference temperature, with ambient as an additional input, best fit the application.

To execute this the following steps were taken for each feature for each image:

1. Masking and import of the Areas of Interest
2. Dimensionality Reduction using simple Descriptive Statistics
3. Comparing the sites using an Least Squares Sum Algorithm
4. Evaluating the Results

4.1.1. Masking and Importing of the Areas of Interest

The first step was the masking of the images to make the extraction as simple as possible. The masking was done in GIMP (GNU Image Manipulation Program). For each of the images a mask

was created for each site. The grayscale image was then imported and for each site the masked area was imported as a full array, preserving the distribution of thermal values in each site. Figure 11 is an example of the masking of an image.

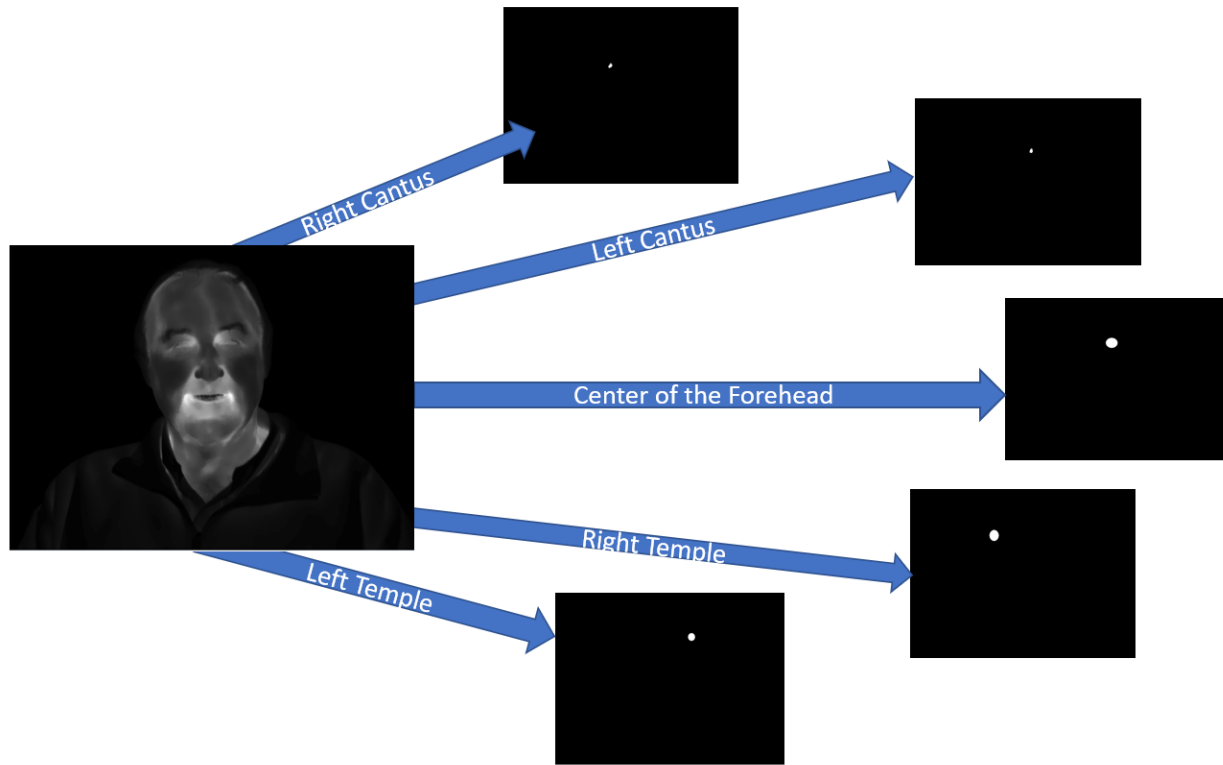


Figure 11 Extraction of Sites

4.1.2. Dimensionality Reduction using simple Descriptive Statistics

For each of the sites, the max, min, mean, median and standard deviation were extracted from the temperature data of the masked areas identified. No advanced dimensionality reduction techniques were used and each site was hand selected. The precision of the mask was subject to human error. The error was reduced by having a single operator utilizing thresholded images with contrast increasement for the warmest regions of the image, however, it was not perfect and the extremely sparse gradient in combination with varying ambient conditions and external factors made it difficult to obtain a perfect mask for each of the images.

4.1.3. Comparing the sites using an Least Square Sum Algorithm

Using these inputs an equation was used to approximate the core body temperature from each site. This equation was found in Houdas and Ring's 1982 publication "Human Body Temperature: Its Measurement and Regulation." The equation can be seen below:

Equation 5 Modified External Heat Transfer Equation [20]

$$T_{core} = T_{site} + ((T_{site} - T_{ambient})^{power} * Step)$$

This equation was slightly modified to incorporate the Step variable, to add precision for the Least Square Sum (LSS) optimization (empirically found to be beneficial). This equation was optimized with a LSS Curve Fit for each site individually. Each possible option for power and step were evaluated in a given resolution to allow for exhaustive evaluation for any potential peaks and valleys in the data. Powers were evaluated from 1 to 9 with a resolution of 1 and the step was between 1E-9 steps from 0 to 1 giving us 100,000 total steps. The difference between the value outputted and the reference body temperature is the error, and this error is squared and summed with the squares of the other errors in the data set. The value that results in the minimal error is saved. This relationship being linear will give a single optimal value that is saved from this LSS fit. Execution of this algorithm with all the steps takes approximately 2 minutes.

The equation could be derived if the previous values were known, however, the data from before or after the image rendering was not obtained. Therefore, this type of exhaustive approach was needed to derive the optimal value, rather than applying more conventional methods for calculation.

4.2. Results

The final step for this experiment was to feed the images' data back through the equation to evaluate if a single site outputted a stable reading with the noted equation.

Figure 12 below shows the loss curves for each of the sites tested:

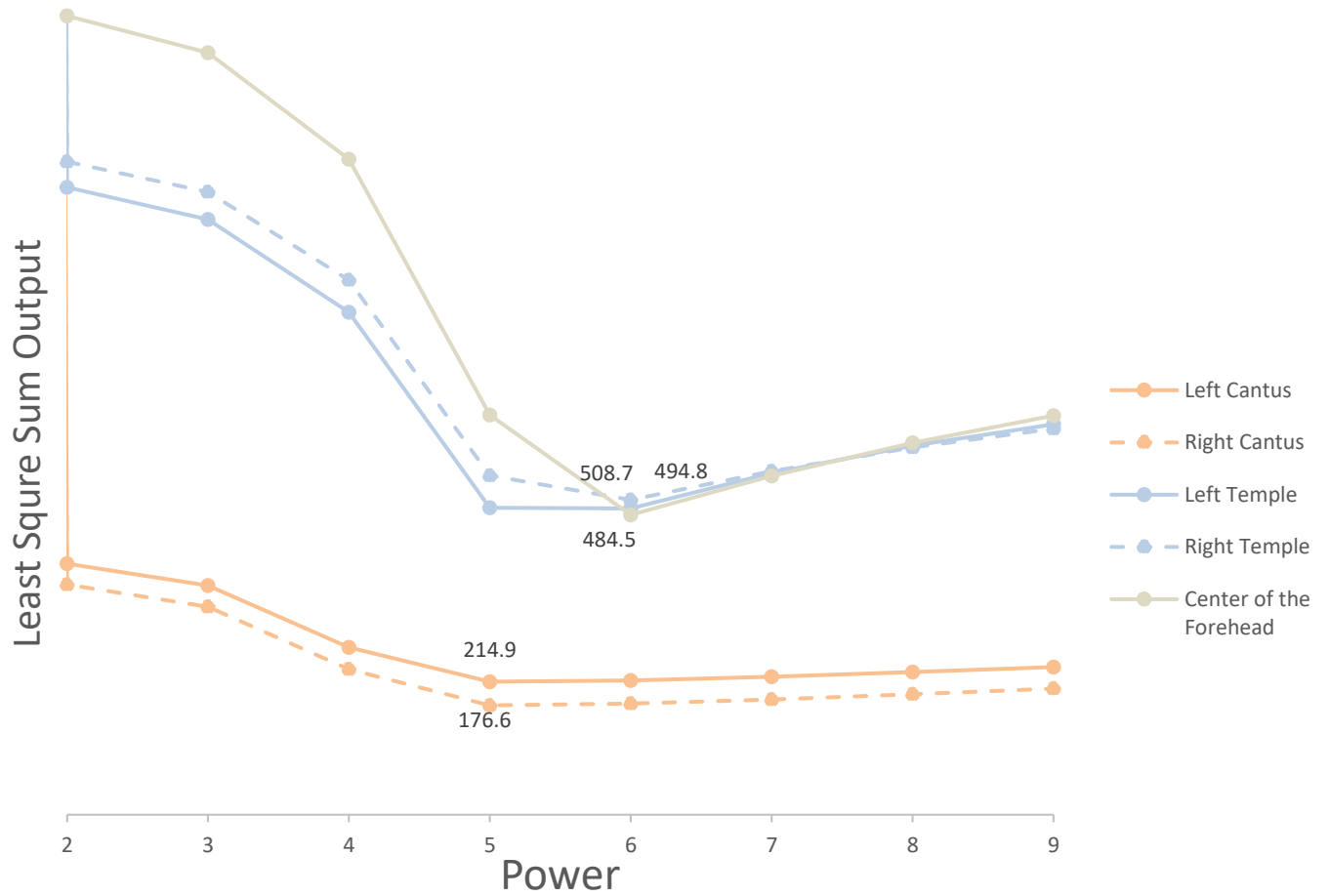


Figure 12 Squared Error verse Power Tested

The mirrored data points (i.e. the canti and the temples) have nearly identical curves. It is also observed that the center of the forehead has similar results as the temples. Table 5 below outlines the equations that were found to be optimal for each site:

Site	Equation
Center of the Forehead	$T_{core} = T_{max} + ((T_{max} - T_{ambient})^6 * (5.72 * 10^{-9}))$
Left Canthus	$T_{core} = T_{max} + ((T_{max} - T_{ambient})^5 * (5.41 * 10^{-8}))$
Right Canthus	$T_{core} = T_{max} + ((T_{max} - T_{ambient})^5 * (5.53 * 10^{-8}))$
Left Temple	$T_{core} = T_{max} + ((T_{max} - T_{ambient})^6 * (5.5 * 10^{-6}))$
Right Temple	$T_{core} = T_{max} + ((T_{max} - T_{ambient})^6 * (5.72 * 10^{-6}))$

Table 5 Equations Outputted by Site

The results of applying these equations to the 55 patients can be seen below:

	Center of the Forehead	Left Canthus	Right Canthus	Left Temple	Right Temple
Bias	-2.2	0.2	0.2	-0.8	-0.8
Standard Deviation	3.3	2.0	1.8	2.5	2.7
Times Site performed the best	15	10	14	8	8

Table 6 Output of Least Square Sum Regression by Site

The results on Table 6 indicate that mirrored sites can be considered analogous and were evaluated jointly. This allows for the analysis of the descriptive statistics to be on a per site basis as seen in Table 7 below. For additional context, descriptive statistics were prepared for a theoretical output (“Best of All”) calculated by selecting the output per image with the lowest error; that is, the center of the forehead for some, the cantus for others, and the temple for the remainder, based on their idiosyncratic performance. This “Best of All” output identifies the

optimal performance achievable with this method as a point of comparison to the performance achieved with any individual site.

	Center of the Forehead	Canthus	Temple	Best of All
Bias	-2.2	0.2	-0.8	-0.2
Standard Deviation	3.3	1.9	2.6	0.9
Times Site performed the best	15	24	16	55

Table 7 Further Results Canthus and Temple Combined

These results demonstrate that the canthus was the best site more often than the other sites, however, it is still not the best site for the majority of the patients. These results are also artificially inflated due to them being evaluated on the same data as they were optimized with. Even with this, high variability is present for each site.

This allows us to draw a conclusion that there is no one best site to be segmented consistently. This is in direct opposition to the earlier research in this area!

An additional step was executed to evaluate the results on only the best of every site. The descriptive statistics can be seen in the 'Best of All' column in Table 7. A Bland Altman Plot was also created to evaluate the results. The Bland Altman Plot is a powerful tool to assess distributions of data where the source of the error is not known, or can be produced from the reference device or the experimental algorithm [4]. The goal for this plot is to have a trend line that is perfectly horizontal and acceptable space between the upper and lower bars. This plot can be seen in the Figure below.

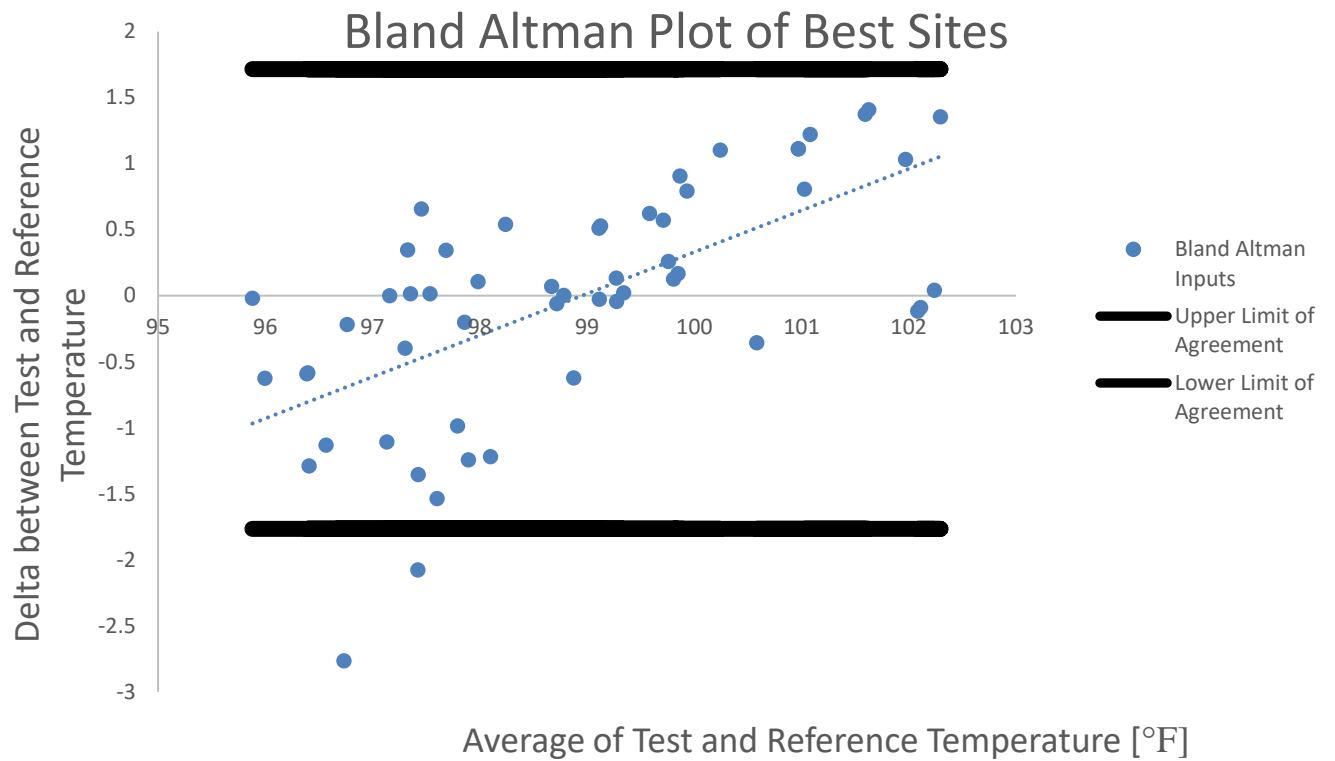


Figure 13 Bland Altman Plot of Best Sites

This demonstrates that the algorithm is predicting high at the high temperatures and low at the low temperatures, but is stable right in the middle of the data. The R^2 value was found to be 0.41, indicating a weak correlation, however, there are two outliers identified in this plot in the 97°F to 98°F range that would strengthen this value if removed, or taken into consideration in a modified algorithm. It is important to reiterate that these results were not evaluated with an independent test set, so they are most likely more accurate than the true error value. Regardless the results still allow the following conclusion to be drawn:

There is no one best site to base a fever screening transfer function from. The relationship is most likely nonlinear in nature and newly developed machine learning techniques can be utilized to characterize this.

5. Binary Classification with Pretrained VGG16

With the presumption that the relationships are nonlinear in nature the deep learning investigation was then explored. Figure 14 Updated Approach with Pretrained Model demonstrates the primary components for this experiment.

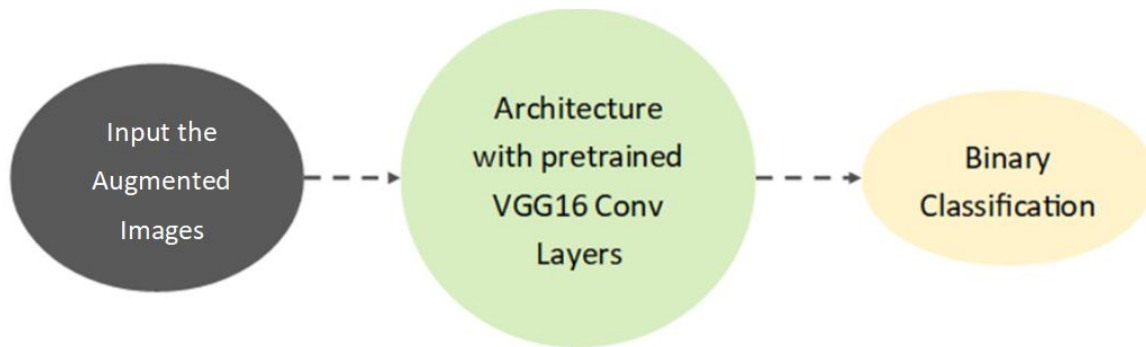


Figure 14 Updated Approach with Pretrained Model

5.1. Implementation

The Pretrained VGG16 Network has many different benefits when compared to an untrained model. These include: developed filters (less data is required to update filters rather than create them from scratch) and starting points for fine tuning of training variables (the hyperparameters that were used in the previous domain are similar to the current). The two important components for this investigation are the handling of the data and preprocessing, and designing the VGG16 used.

5.1.1. Dataset and Preprocessing

The data was selectively limited by incorporating a single criteria. The only requirement was for both eyes to be present in the image evaluated. This reduced the dataset down to 160 images total and 35 of those images contained febrile individuals.

To maximize the benefit of using the pretrained model it was found to be beneficial to modify the input images' dimensions to the scale of the images in the original domain of the network (scaling the maximum value to 255). It was also found to be beneficial to decrease the resolution

of the images from 640x480 to 320x240. This allows for the full face of the individual to be incorporated in the crop, which was not the case with the larger images, allowing the network to identify more common features. Due to deep learning's dependency for data to train the values against, image augmentation was used to increase the number in the dataset with more legitimate data. These images of size 320x240 were then cropped to 224x224 with a structured crop to result in 5 differently cropped versions of each input image, and then the augmented inputs were flipped to double the amount of individuals present. As the pretrained network requires images with RGB channels, the black and white image data was duplicated 2 times and appended to the original to give 3 identical channels, giving 10 augmented inputs with dimensions of 224x224x3 for each image in the dataset. This technique was found superior to images that were randomly cropped, due the systematic approach. This allows for all features to be present in the augmented images rather than potentially removing features due to the random placement of the crop on these images.

The structured crop was designed to displace the individual in the image while taking advantage of the convolutional property of location invariance. The structured crop was conducted by rendering a square over the center of the nose of the patient in the image. The augmented inputs were cropped from the corners and center of this square; as the data required both eyes to be present in the base image there would always contain part of the sides of the face present in the resultant input using this technique. The output of this type of crop is the shifting of an individual from the center of the image to the four corners of the image. The final results are using only the base output of the network and not using any additional processing (such as a vote method explained in Chapter 7). This was the final design for the traditional convolutional architecture to allow for maximum effectiveness and is the same method used in Chapter 6. Examples of this means of augmentation can be seen in the figure below.



Figure 15 Structured Crop Example VGG16

5.1.2. VGG16 Network

This augmented data was then passed through a pretrained VGG16 network. This pretrained model was trained from the ILSVRC (ImageNet Challenge) denoted in section 1.6. The dataset that this network was trained on was composed of millions of images classifying one thousand different classifications. For this research, the convolutional layers were left untrained, leaving only the fully connected layers and the Softmax classification layer trainable. The fully connected layers, while pretrained, were trainable so that the important features could be learned (denoted

in the gray box seen in the figure below). This is outlined in

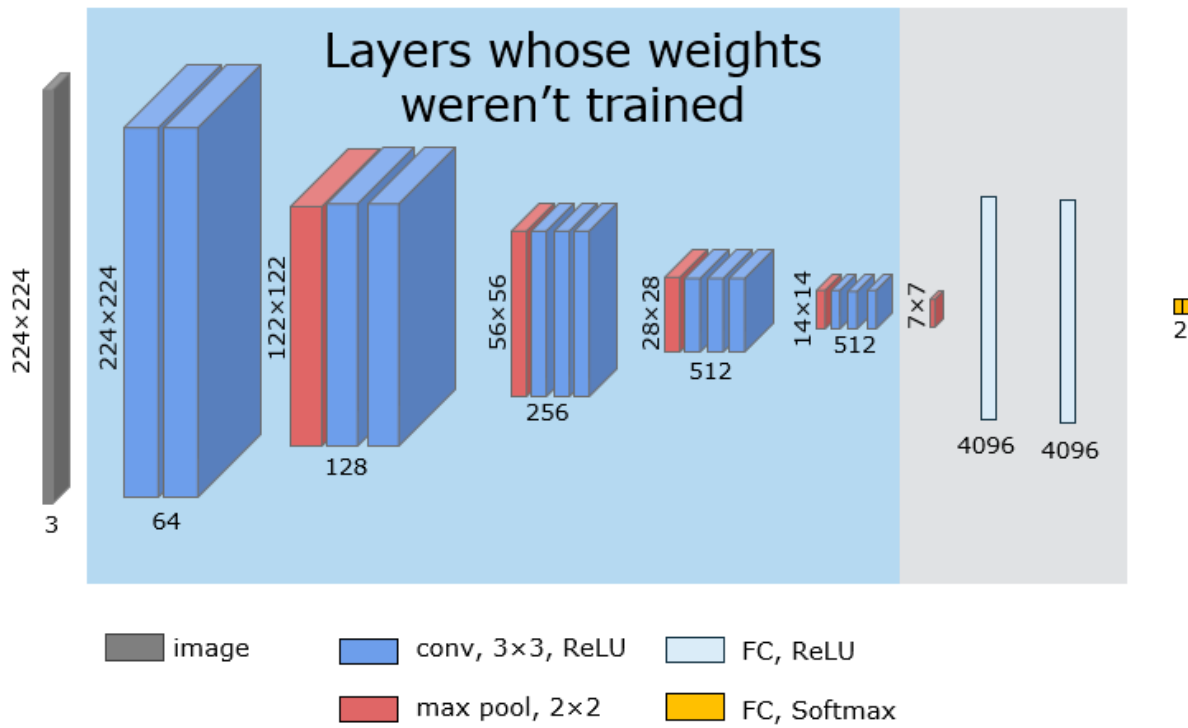


Figure 16 VGG16 *Architecture*:

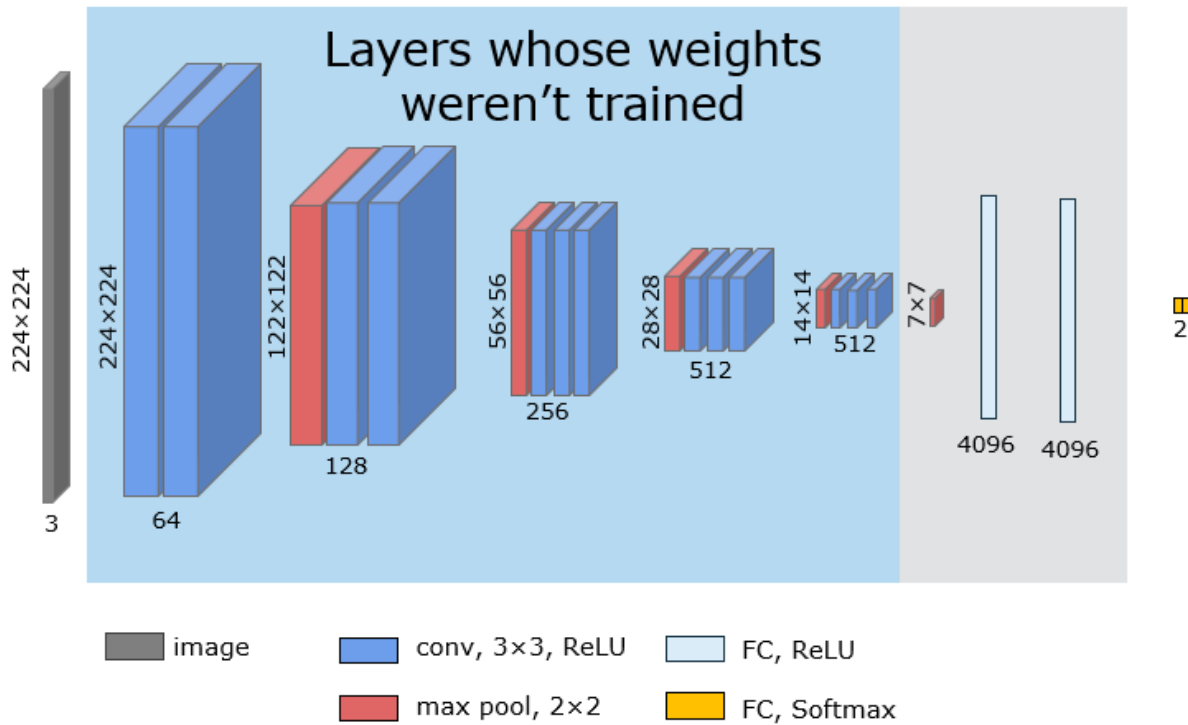


Figure 16 VGG16 Architecture

5.2. Experiment Results

The results of this method are outlined in Table 8 Pretrained VGG16 Results with the Full Face and Structured Crop:

Accuracy	84.7%
Sensitivity	45.7%
Specificity	96.0%
Prevalence	22.4%
F1	0.57

Table 8 Pretrained VGG16 Results with the Full Face and Structured Crop

This table outlines a large discrepancy in the sensitivity and specificity. This is due to the network being imbalanced, with more afebrile data present than febrile. Due to the network being probabilistic in nature, the network will learn that guessing "afebrile" would result in less loss on average than guessing febrile; the probability of being wrong when guessing afebrile is very

low simply due to the probability of the subject being febrile is very low. In order to balance the sensitivity and F1 score corrective actions in the design of the network must be taken.

5.2.1. Balanced Network Results

There were a few options for correction of an imbalance dataset. These options are weighted binary cross entropy (modifying the base cross entropy equation to give more of a penalty for incorrect classifications of fever), up-sampling of febrile data (artificially increasing the number of febrile images in the dataset) or random down sampling of the afebrile samples of the dataset (randomly remove base images from the training set).

Up-sampling the febrile data is a technique with risk involved. It allows for any atypical, detrimental data in the training set to have an unjustifiably large impact on the effectiveness of the training and to artificially decrease the accuracy of the technique. An issue with augmentation being an untrue representation was found in the initial implementations of the experiment, and it was believed to have a high probability of reoccurring if up-sampling was introduced. For this reason, weighted binary cross entropy and random down sampling of the afebrile samples of the dataset was used to balance the prevalence of fever in the dataset.

Weighted binary cross entropy is the modification of the weights seen in the equation outlined on page 23 to accommodate decreased prevalence. The values found to give the best balance were 0.8 for febrile and 0.2 for afebrile data (from 0.5 for both febrile and afebrile data), and the final equation for this can be seen below:

Equation 6 Weighted Binary Cross Entropy Equation

$$loss = -(0.8 * output * \log(probability) + 0.2 * (1 - output) * (\log(1 - probability)))$$

The semi-randomly (afebrile only) down sampling of the dataset was executed by applying a random value to all data that was found less than 99.5°F (37.5°C), sorting the randomly applied values in order from greatest to least and selecting 35 samples with the lowest randomly applied values.

By applying weighted binary cross entropy and a semi-randomly down sampled dataset, the following results were obtained on a single trial:

Accuracy	78.7%
Sensitivity	87.1%
Specificity	70.3%
Prevalence	50.0%
F1	0.80

Table 9 Results of the First Down Sampled Network

When comparing these results to the Base Pretrained VGG16 sample, the data demonstrates the decreased tendency for classifying all inputs as afebrile via the decreased discrepancy between the Sensitivity and the Specificity (the network classifies febrile more often with these improvements). This can be seen in Figure 17 Slope Graph Comparing the Base Pretrained VGG16 Network with the Semi-Randomly Down Sampled VGG16 Network.

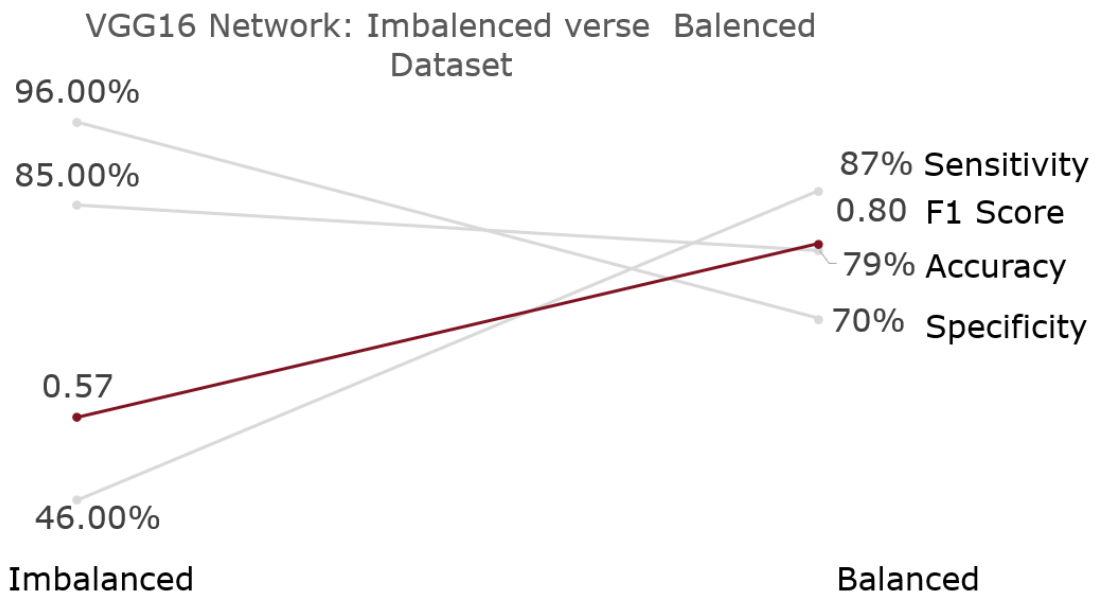


Figure 17 Slope Graph Comparing the Base Pretrained VGG16 Network with the Semi-Randomly Down Sampled VGG16 Network

5.3.Conclusion

Over all, this algorithm does display acceptable results that meet the hypothesis of obtaining a F1 score of 0.80. This model is realistic based upon the noisy data gathered and demonstrates that this algorithm is promising if the used with a large dataset or if used with less noise intentionally added in the gathering of the data. This experiment made it clear that pretraining could not be the only answer to improve the algorithm with data limitations. The results found in this investigation met the primary objective, but only barely, with a single down sampled dataset. Another technique had to be utilized to accomplish the primary objective..

6. VGG16 Feature Extractor to a Principle Components Analysis to Support Vector Machine Approach

The fully connected layers of the Neural Network, which were the primary decision-making layer of the previous classification, were not the most efficient means of classification for this research due to the data limitations. Principle Components Analysis (PCA) with the aid of a Support Vector Machine was found to be a more effective way to classify fever status with minimal data. Below is the updated design to accomplish this algorithm.

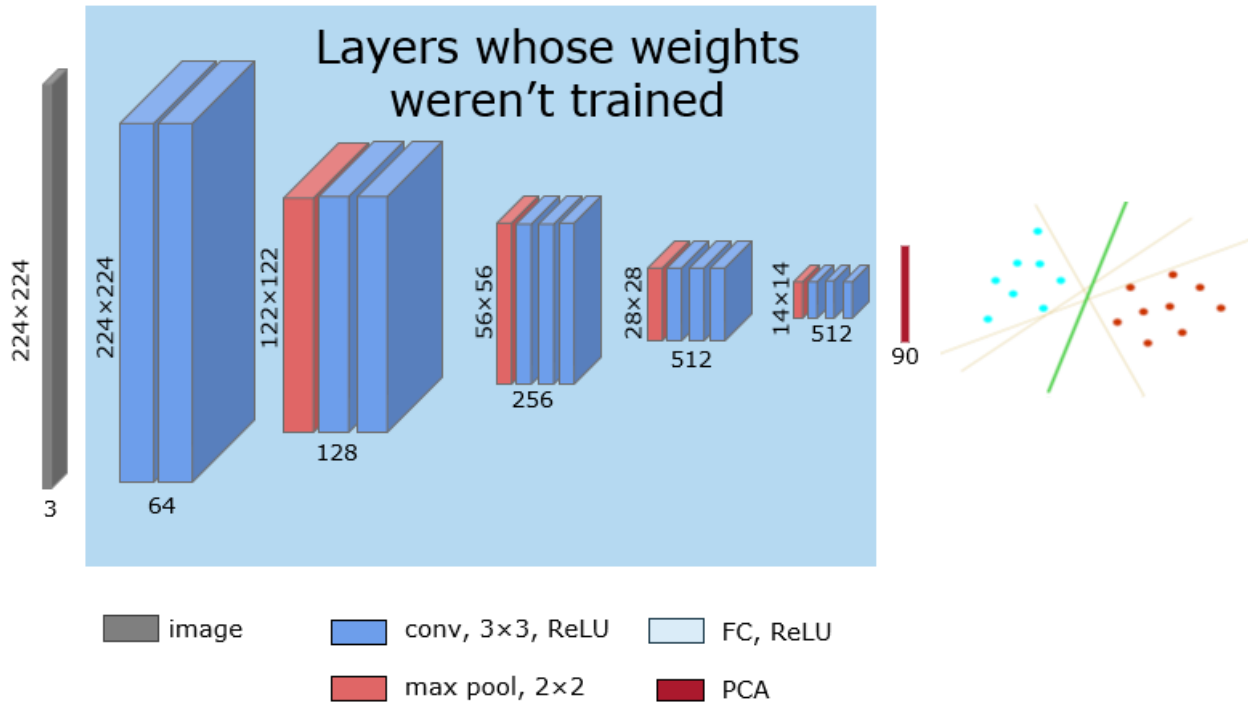


Figure 18 Final Approach for Classification

6.1. Execution

The new technique was designed to make the fully connected layers obsolete due to their dependence on data for training (this was not the correct answer of this dataset). The goal for

the classification was to design to utilize the intrinsic relationships outputted by the convolutional layers.

6.1.1. Preprocessing and Augmentation

The data used in this experiment utilized the structured crop, however the data was not flipped around the y axis due to the technique's decreased dependency for data. An example of a cropped image can be seen in the figure below:

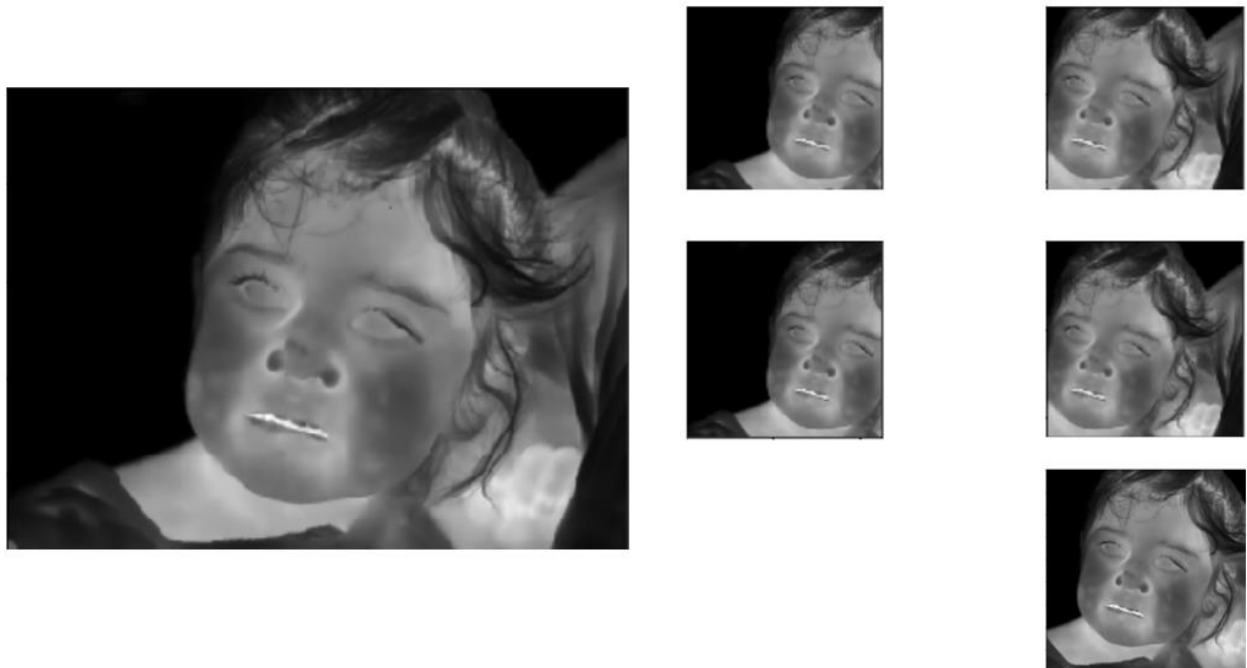


Figure 19 Structure Crop with Increased Contrast

In addition to the crop an additional preprocessing technique was used. Each pixel of the images (prior to segmentation) was squared, then normalized to 255, with the max value for each image being 255 and all other points being their square relation to the new squared max value. The normalization to 255 allowed for maximum overlap between the dataset and the domain the preprocessed feature extractor was trained on.

The theory behind this preprocessing is mechanical in nature. Due to the high precision of the measurement equipment (as indicated in Section 2.2 Thermal Imager), and the ambient temperatures found in the investigation being lower than the physiological temperature present,

the assumption can be made that the warmest regions in the image were physiological structures and of interest in the classification. Squaring these points helps push the warmest regions 'forward,' giving them more weight for the deep learning algorithm. It was empirically found that higher orders of magnitude (i.e. cubing or putting the data to the 5th power) had a detrimental effect on the results. It is hypothesized this is because it pushes the secondary sources of heat further back, giving all additional weight to the warmest feature or draws too much attention for too few pixels in the area of interest (only a part of the warmest region could be used due to a large discrepancy in the feature's thermal characteristic).

An image demonstrating what increasing the data by a power, then normalizing the scale can be seen in the figure below. The 10th power was used in the image to enhance the contrast between the two images (to make it distinguishable for humans), however, the 2nd power was found to be optimal (empirically).

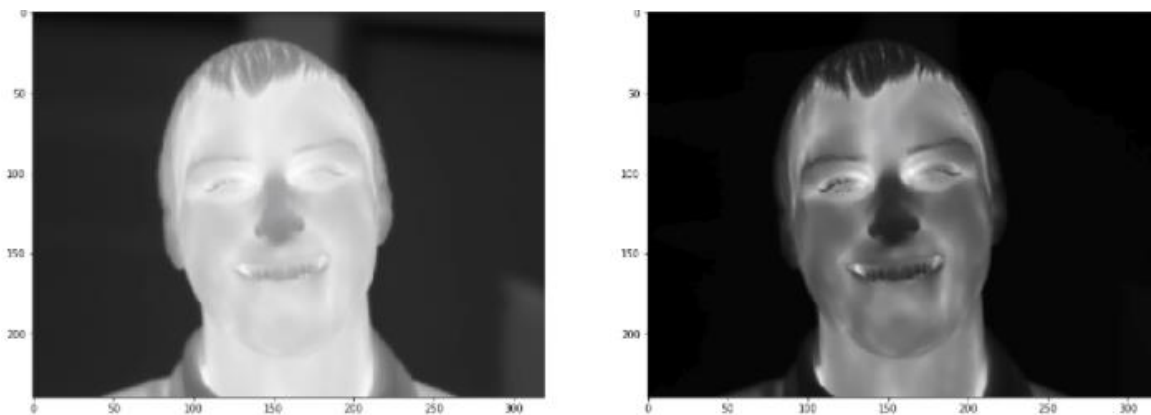


Figure 20 Base image (left) and Image raised to 10th power and scaled(right)

The graph below demonstrates that the second power outputted the best results of the powers tested. The maximum accuracy was found to peak at this second power. This was conducted with exhaustive computation of various orders of magnitude. Due to the computational efficiency of the technique, empirical optimization started with putting each point in the data to the 5th power, then the 3rd and finally finding superior results at the 2nd power. This was then evaluated with 2 different down sample trials where the results were confirmed. The figure below outlines the initial trial's results.

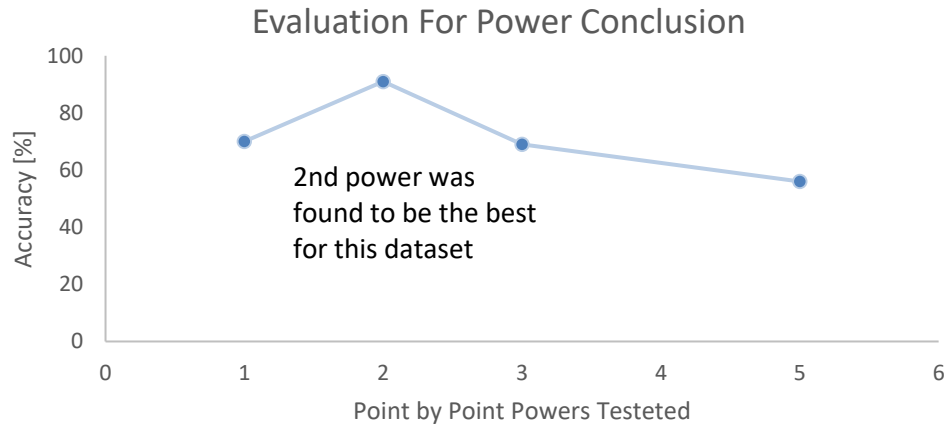


Figure 21 Evaluation for Point by Point Power

6.1.2. Feature Extraction

To utilize the strength of all components used, the same pretrained VGG16 Convolutional Neural Network was used to identify important nonlinear relationships in the inputs, commonly known as a feature extractor. Using the end output of the final Convolutional layer in this pretrained model allows for various, defined blocks of data that simplify nonlinear patterns intrinsic to the inputs, just as they do in the VGG16 Model. These layers were left untrained to not disrupt the defined filters with this research's lack of data. This helped further decrease the sample size limitation.

At a simplistic core component level, the Convolutional Neural Network is made up of multiple small images for each of the original 350 input images made by passing the pixels of the original input images through multiple filters. Each small filters has dimensions of 14x14x512 data points; the specific structures in these small images are targeted in an attempt to simplify nonlinear trends in the data.

Principle Components Analysis (PCA) was chosen to help decrease the dimensions of the data and quantify the information in a more meaningful manner. This technique was used because it is hypothesized that the noise in the dataset was intrinsic to the fully connected layers. If the noise also came from the convolutional layers, the feature extractor would not be successful.

However, if the problem was lack of data for fine tuning the fully connected layer, the PCA-SVM model should give superior results due to less data dependency in the technique.

6.1.3. Principle Components Analysis

The power in this investigation lies in the nature of PCA. PCA is a technique used by many data scientists over the past decade to reduce a large amount of data points of a common feature to a single point. This is known as dimensionality reduction as each point of data adds a new dimension and requires analysis. This means that the number of important features inside the data can be reduced into a few meaningful ones. For this application, the feature extractor outputs a vector of 100,352 data points ($14 \times 14 \times 512$ points) for each input. This data then gets categorized into 90 principle components (90 was empirically derived).

This exhaustive model to find the characterization of the components was possible due to the computational efficiency of the technique. PCA's end goal is to condense the data points of a related feature to a single base vector where the presumed linearity of the data lies (multiplied by a dot product) [46]. It is assumed that the data can follow a linear trend, i.e. if plotted, the data would appear elliptical in nature rather than circular. The elliptical would be more useful because it would be capable of converging to a centered orientation rather than any potential distribution. Figure 22 illustrates the difference between a distribution that allows for a single center vector (the ellipse) versus a distribution that allows for multiple center vectors (the circle).

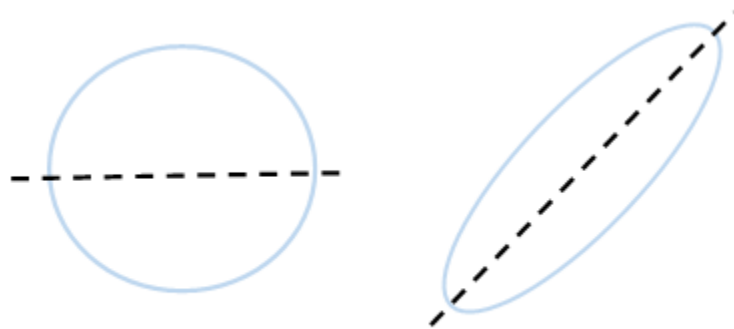


Figure 22 Circular Distribution versus Elliptical Distribution with vector

This figure also demonstrates the variability present in the input vectors; the difference between any given point and the center vector is referred to as variance. The variance of the data is defined as the sum of the squared distances from each input vector to the center vector. Components (or areas) with high variance are overly influenced by (or represent or are caused by) noise in the data [46].

This algorithm finds similar clusters of data points outputted by the feature extractor by comparing the local subsets of the points to sets of orthogonal vectors. This multitude of data points can be reduced to a handful of meaningful points to be fed into the Support Vector Machine (SVM).

6.1.4. Support Vector Machine

A traditional support vector machine was found to give the best results. The SVM takes the data fed from the PCA and considers the output a part of a hyperplane, a multidimensional plane, to simplify the data to a single classification entity. As the name implies, the algorithm does this by applying vectors to separate the data (essentially multidimensional thresholds).

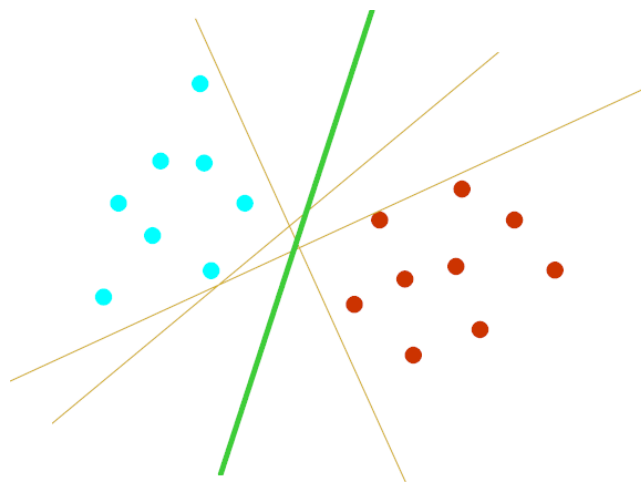


Figure 23 Optimal Separating Hyperplane [14]

The algorithm optimizes the placement of the multidimensional in the hyperplane by minimizing the error of a training set by optimization of an equation, linear or nonlinear in nature, that defines the separating vectors. For our investigation, it was found that a third-degree

polynomial kernel with a kernel coefficient of 0.0111 (1/ 90, 90 chosen due to it being the number of components used in the PCA) outputted the best results (found empirically).

6.2.Results

6.2.1. Base Results

The initial results were just the product of the output directly from the SVM. The algorithm was run 5 times due to the down sampling of the of the images selected for the input. The average output at the SVM can be seen below:

Accuracy	83%
Sensitivity	74%
Specificity	93%
F1	0.82

Table 10 PCA-SVM Results

After analyzing the heatmap of the various trial results it was evident that there was noise in the output (seemingly random incorrect classifications of data throughout the dataset). Figure 24 below is this heatmap, where the green values are the correct classification and the red are incorrect classifications.

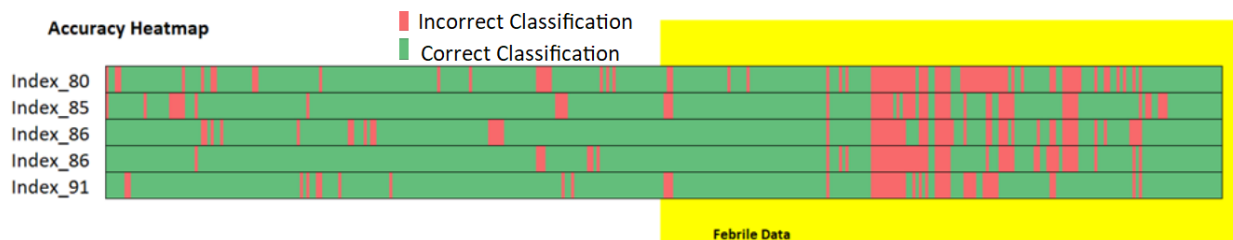


Figure 24 Accuracy Heatmap of SVM-PCA Approach

6.2.2. VOTE

To combat the random noise in the output of the data, the results began to be analyzed by the base (non-augmented) image that was augmented rather than the augmented inputs. This was done by considering all results from the augmented inputs of a given image as subcomponents of the same image. If the majority (3 or more) were predicted febrile, then the full input image was found to be febrile, and vice versa. The data was then reanalyzed for each randomly down sampled index.

This resulted in slight improvements to the Sensitivity of the network at the slight expense of the Specificity. The improvement in the Sensitivity is greater than the loss of the Specificity, resulting in an increased the F1 Score. These results can be seen below:

Accuracy	85%
Sensitivity	80%
Specificity	91%
F1	0.84

Table 11 Feature Extractor - PCA - SVM - Vote Average Results

Less noise was found in the heatmap, and the reduction in noise made outlier detection less difficult. This was executed by analyzing each image for each trial (each different randomly down sampled index). The seemingly random variation in the results is reduced with this technique, making it easier to identify outliers. If a single image was found to have inaccurate readings for the majority of the trials (a misclassification for a minimum of 3 of the 5 trials) it was identified as an outlier. The Heatmap below outlines the accuracy heatmap of the PCA-SVM-Vote, and the blue bars underneath identify the outlier readings.

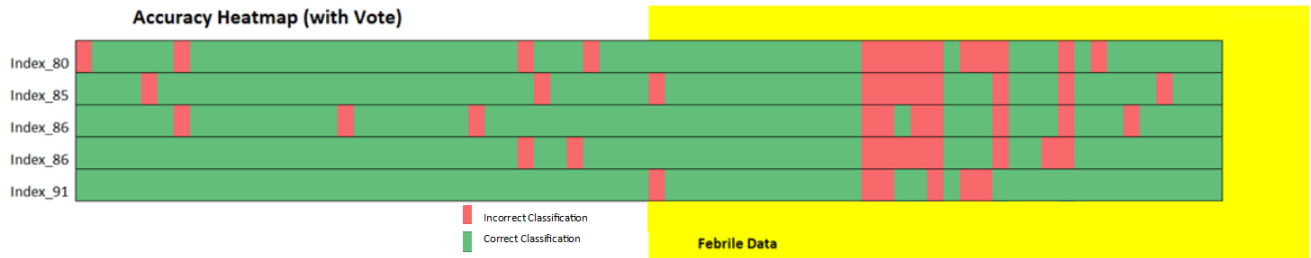


Figure 25 PCA-SVM-Vote Accuracy Heatmap

7 images were visually inspected to see if any abnormality could be the root cause for their misclassification. These images can be seen below:

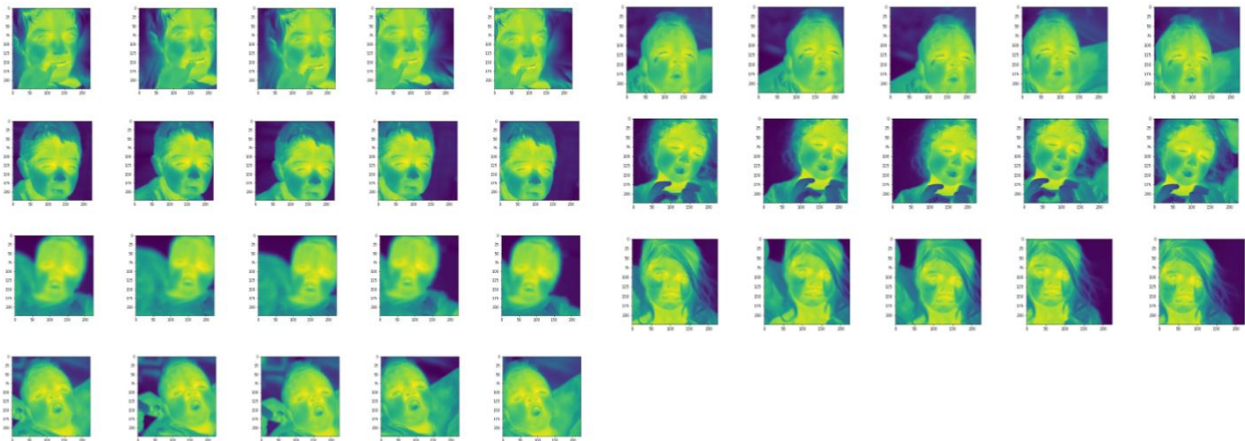


Figure 26 Outliers Identified in PCA-SVM-Vote! Results

From these images there were a couple patterns identified. First, from the heatmap it is evident that all images were in the febrile class. Not only are they in the febrile classification, but they were scattered across the febrile region (core temperatures from 99.7°F [37.6°C] to 101.3°F [38.5°C]). In addition to this, they are all within the requirements of the data gathered.

Two relevant similarities were identified. First is that they are all children from the same clinical site, and second, one child was an outlier three times (for all the images present in the dataset). Due to a lack of disqualifying data or any apparent discrepancies in the inputs, none of the identified outliers were filtered and all data was included in the final results for this algorithm.

6.3. Experiment Conclusion

The following results demonstrate that this is an extremely viable method for the classification of febrile subjects in the data. The final step of this research is to compare the two feasible methods and evaluate if one is superior.

7. Final Experiment: Comparison the Two Methods

7.1.Design

With two viable algorithms that can potentially accomplish the objective of the hypothesis, the final experiment conducted was a comparison of the results Semi-Randomly Down Sampled Pretrained VGG16 Network and the VGG16 Feature Extractor to the PCA to the SVM to a Vote (from here further it will be referred to as the Test Algorithm). The first action was to run both algorithms 5 times, which is needed due to the semi-random down sampling. The results need to confirm that there is a difference between the VGG16 output and the Test Algorithm's output, independent of sample index.

Once 5 trials were conducted, the next step was to apply a Lilliefors test to the F1 Score and Accuracy of the results. The Lilliefors test is a statistical method that evaluates the normality of the outputs of the data [29]. If the Lilliefors test demonstrated that the distributions of the resulting performance metrics were normal than a T test would be used to evaluate if the means of the distributions were the same. If the distribution was found to not be normal then a Wilcoxon Ranked Sums test will be used instead. The difference of the distributions would be evaluated at 95% confidence.

7.2.Results

The Lilliefors test for normality confirmed that the data could be treated as normal (more specifically, it failed to conclude that the data was not normal). This is due to the p values for the VGG16 results were 0.3681 for the F1 score, and 0.4841 for the accuracy. The Test Method (Feature Extractor to the PCA, SVM and Vote method) results were 0.2035 for the F1 score, and 0.2796 for the accuracy of this design.

Due to the inability to reject normality for the distributions, a T test was utilized to evaluate if the distributions that were statistically different. The p value for the F1 values was computed to be 0.014 and the p value for the accuracy was found to be 0.01.

These results demonstrate with 95% confidence that the outputs of the algorithms are not the same, and due to the higher performance metrics, the Test Algorithm is found to have significant better performance compared to the VGG16 Network for this dataset.

The results of the two networks can be seen on the figure below:

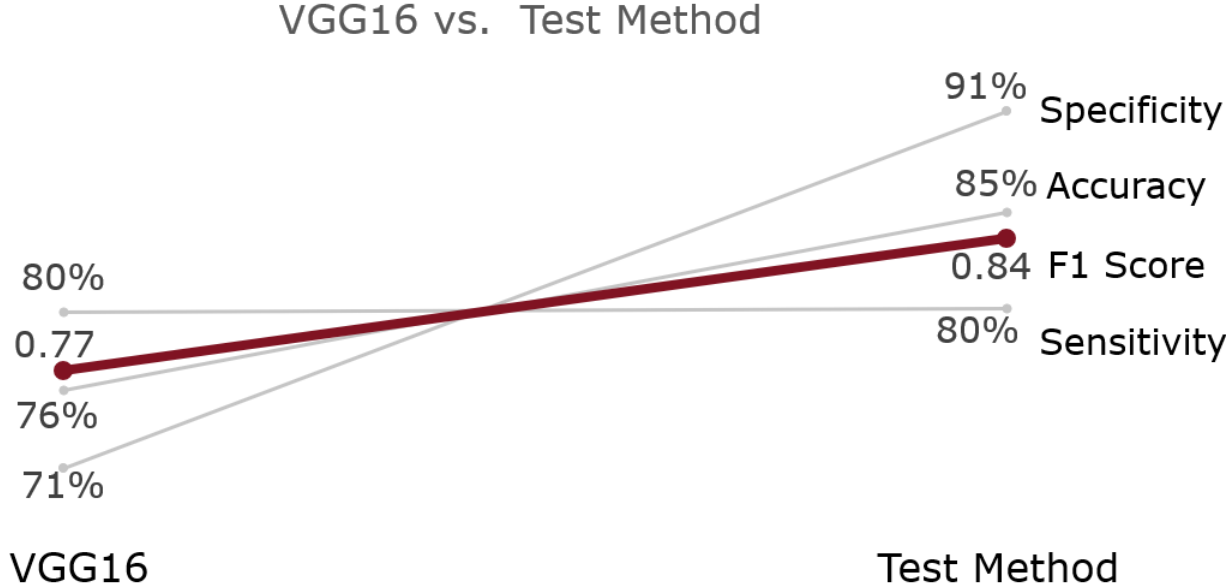


Figure 27 Comparison of the Pretrained VGG16 with the PCA-SVM-Vote! Algorithm

7.3.Conclusion

The results from this final network meet the requirements of the research hypothesis validating that Machine Learning techniques can handle this noisy data and correctly classifying to the appropriate febrile status.

8. Discussion

8.1. Primary Scope

The primary goal for this investigation was to evaluate the viability for modern Machine Learning techniques to identify febrile status in the presence of noise. This noise includes distance, emotional status, facial orientation, ambient temperatures and acclimation times. Trying to tackle this problem in the presence of noise is an EXTREMELY difficult task.

The first and one of the more common obstacles experienced, especially with proof of concept research, was number of samples present. Sample size used in training is a luxury that this research was unable to afford. The deep learning approach is a big data technique, typically created using large sample sets. Various experts in the field claim that the technique should not be attempted with sample sizes under 1000. The inability to approach these sample sizes was the primary motivation to use the CNN strictly as a feature extractor that did not require prior training. The final validation of the network was conducted on 70 samples for any given trial without sufficient data to allow stratification for various categories of individuals (acclimation time, age, core body temperature, etc). This validation is sufficient for proof of concept but is not sufficient for the final validation of an algorithm.

The other studies outlined in 1.5 Infrared Measurements in Practice limited the noise by controlling the external variables, including dictating specific measurement sites. Ng 2012 really began to advance by combining two sites with a specific algorithm meant to gauge multiple variables. This research rejects both models (to an extent) due to the development of machine learning over the past decade. The year Ng published his work, AlexNet revolutionized computer vision using convolutional neural networks for nonlinear transformations for object recognition. Over the past decade there has been remarkable progress in these networks. To the researchers' knowledge, this is the first application of CNNs for the classification of febrile patients, and it is with the added difficulty of the noise caused by real world external variables present in the data. The true goal of these experiments was to evaluate the feasibility of classifying in the presence of noise in an uncontrolled setting due to the previous publications demonstrating the feasibility of correct fever screening in controlled settings.

This noise is an extremely necessary obstacle to overcome. The author's experience with consumer goods shapes the opinion that all devices and algorithms, regardless of how elegantly designed, are susceptible to user error, and that error will be experienced. For example, this thermal scanning device may be placed near the entrance of a building to prevent an ill person getting too far into the premises. The device would experience wide fluctuations in ambient temperature which may unintentionally force an incorrect result, with no intention of doing so. The results previously outlined in this thesis **do not** demonstrate a full successful algorithm, ready for consumer use, but they do demonstrate that it is possible to overcome these obstacles using improved processing methods, rather than artificial constraints on the measurement variables.

The impact of the external variables (manifesting itself as noise in the images) cannot be overstated. Measurement in the presence of this 'noise', rather than artificially constraining the external variables that cause the 'noise,' directed the route for the approaches taken. It is conventional knowledge in noninvasive thermometry to target the specific sources of heat. The acclimation time directly impacts the route chosen for this. When the individual is not acclimated to their environment there is too sparse of a gradient to find a distinguishable source of heat. A resulting image of such an individual can be seen below:

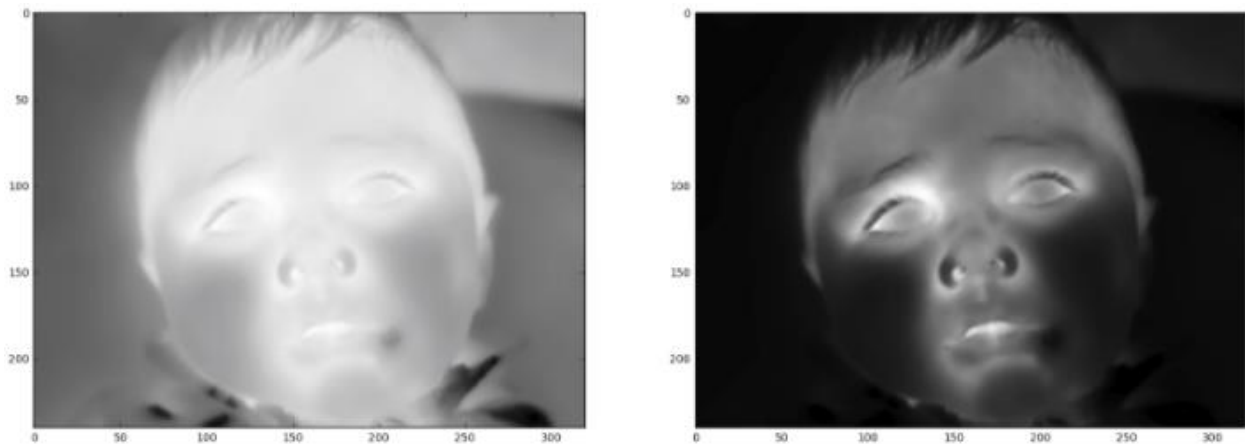


Figure 28 Unacclimated Patient

This is why the initial investigation (found in Chapter 4) was conducted using objects that could be identified via digital image. Once this investigation demonstrated the difficulty in the

secondary investigation it was clear that a more complex algorithm would be required to characterize the nonlinearity.

This research proved to be a viable route. The imbalanced dataset immediately impacted the results, giving a high specificity and low sensitivity. Various different forms of preprocessing (histogram processing, targeted contrast enhancement, k-means algorithms with various centroids, etc.) were investigated to try and increase the accuracy, however, nothing helped these techniques with the minimal amount of data collected. Conventional methods of balancing the dataset while using unaltered images proved to be the best method. It became evident that the Fully Connected Layer was not the best technique for our application and the Feature Extractor-PCA-SVM-Vote was conceived.

The Feature Extractor-PCA-SVM-Vote technique brought the best of all the worlds together; it required less data to make the classification while still extracting the nonlinearity in the data. This gave the anticipated results of improvement to the F1 score. The neural networks are extremely efficient at identification of nonlinearities in the data once trained, however, if there are not enough data to train these fully connected layers, then the results will not be sufficient. Therefore, it is hypothesized that if more training data was introduced to the dataset the result would be increased performance for the VGG16.

In all, the Feature Extractor-PCA-SVM-Vote method results meet the objective of the hypothesis and demonstrates that it is possible for an algorithm to overcome these obstacles. Potential areas that this research can be expanded upon are, increasing the F1 score and accuracy of the model, quantifying how human acclimation from the environment and objectively measuring which facial feature is the best to classify febrile status.

A. Bibliography

1. Amalu, William C., Jonathan F. Head and Robert L. Elliot. "Infrared Imaging of the Breast: A Review." *Medical Infrared Imaging Principles and Practices*. By William B. Hobbins. Boca Raton, FL: CRC, 2013. Print.
2. Bhowmik, M., S. Kankan, S. Majumder, G. Majumder, A. Saha, A.N. Sarma, D. Bhattacharjee, D.K. Basu and M. Nasipuri. "Thermal infrared face recognition—a biometric identification technique for robust security system." *Reviews, refinements and new ideas in face recognition* (2011): 113-138.
3. Bitar D., A. Goubar, J.C Desenclos, "International travels and fever screening during epidemics: a literature review on the effectiveness and potential use of non-contact infrared thermometers", *Euro surveillance* (2009).
4. Bland, J. M., and D.G. Altman. "Statistical Methods For Assessing Agreement Between Two Methods Of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307–310., doi:10.1016/s0140-6736(86)90837-8.
5. Blumberg, Mark S. "Body Heat", *Harvard University Press*, 2009. ProQuest Ebook Central, <http://ebookcentral.proquest.com.ezproxy.wpi.edu/lib/wpi/detail.action?docID=3300584>.
6. Cao, L.j., and W.k. Chong. "Feature Extraction in Support Vector Machine: a Comparison of PCA, XPCA and ICA." *Proceedings of the 9th International Conference on Neural Information Processing*, 2002. ICONIP '02., doi:10.1109/iconip.2002.1198211.
7. "Celsius." *The New Dictionary of Cultural Literacy: What Every American Needs to Know*, edited by E. D. Hirsch. 3rd ed., *Houghton Mifflin*, 2002. Academic OneFile
8. Chan L.S., G.T. Cheung, I.J. Lauder and C.R. Kumana. "Screening for fever by remote-sensing infrared thermographic camera." *J Travel Med*. 2004;11(5):273-9. 12.
9. Chiu W.T., P.W. Lin, H.Y. Chiou, W.S. Lee, C.N. Lee, and Y.Y. Yang, "Infrared thermography to mass-screen suspected SARS patients with fever." *Asia Pac J Public Health*. 2005;17(1):26-8.
10. Cisneros, Austin B., and Bryan L. Goins. "Body Temperature Regulation". *Nova Science Publishers, Inc.*, 2009. ProQuest Ebook Central,
11. E. F. J. Ring, "Progress in the measurement of human body temperature," *IEEE Engineering in Medicine and Biology Magazine*, vol. 17, no. 4, pp. 19-24, July-Aug. 1998. doi: 10.1109/51.687959 <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=687959&isnumber=15099>
12. Goodfellow, Ian, Yoshua Bengio and Aaron Courville. "Deep Learning." *MIT Press*, 2017. www.deeplearningbook.org

13. Greenes DS, G.R. Fleisher. "Accuracy and tolerability of a non-invasive temporal artery thermometer for use in infants." *Pediatric Academic Societies/American Academy of Pediatrics Conference*, 5/2000.
14. Gunn, Steve R. "Support Vector Machines for Classification and Regression." *University of Southampton*, 1998
15. Hausfater P, Y. Zhao, S. Defrenne, P. Bonnet, B. and Riou "Cutaneous infrared thermometry for detecting febrile patients." *Emerg Infect Dis*. 2008;14(8):1255-8.
16. Herzog, Lynn, and Stephanie G. Phillips. "Addressing Concerns About Fever." *Clinical Pediatrics*, vol. 50, no. 5, 2010, pp. 383–390., doi:10.1177/0009922810385929
17. Hinton, Geoffrey, Nitish Srivastava, and Kevin Swersky. "Lecture 1c Some Simple Models of Neurons." *Neural Networks for Machine Learning*. Toronto, Canada, *University of Toronto*.
18. Hinton, Geoffrey, Nitish Srivastava, and Kevin Swersky. "Lecture 3c The Backpropagation Algorithm." *Neural Networks for Machine Learning*. Toronto, Canada, *University of Toronto*.
19. Hogan, David E., S. Shipman and K. Smith "Simple Infrared Thermometry in Fever Detection: Consideration in Mass Fever Screening." *American Journal of Disaster Medicine*, www.wmpllc.org/ojs-2.4.2/index.php/ajdm/article/view/214.
20. Houdas, Y., and E. F.J. Ring. "Human Body Temperature: Its Measurement and Regulation." *Plenum Press*, New York and London, 1982. Print.
21. "Influenza (Flu)." *Centers for Disease Control and Prevention*, 2 Nov. 2017, www.cdc.gov/flu/pandemic-resources/basics/past-pandemics.html.
22. "Infrared Sensor Stabilizable in Temperature, and Infrared Thermometer with a Sensor of This Type." Bernhard Kraus, assignee. Patent 6626835B1. 30 Sept. 2003. Print.
23. Kaiser, Manfred. "Radiation Thermometer and Method of Computing the Temperature." Bernhard Kraus, assignee. Patent 6149298A. 21 Nov. 2000. Print.
24. Kingma, Diederik P and Jimmy Lei Ba. "ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION." *ICLR*, 2015, arxiv.org/pdf/1412.6980.pdf.
25. Kim, Seunghyeon, Wooyoung Kim, Yung-Kyun Noh and Frank C. Park, "Transfer Learning for Automated Optical Inspection." *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, doi:10.1109/ijcnn.2017.7966162.
26. Klos, Alexander, Elke Kahler, Frank Beerwerth, and Horst Mannebach. "Infrared Thermometer with Heatable Probe Tip and Protective Cover." Bernhard Kraus, assignee. Patent 6694174B2. Jan.-Feb. 2004. Print.
27. Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.
28. LeCun, Y., L. Bottou, Y. Bengio, P. and Haffner. "Gradient-Based Learning Applied to Document Recognition ." *IEEE Explore, IEEE Journals & Magazine*, Nov. 1998, ieeexplore.ieee.org/document/726791/.
29. "Lilliefors Test." *Oxford Reference*, 3rd ed., Oxford University Press, 2014.

30. Liu C.C., R.E. Chang, W.C. Chang. "Limitations of forehead infrared body temperature detection for fever screening for severe acute respiratory syndrome." *Infect Control Hosp Epidemiol.* 2004;25(12):1109-11. 15.
31. Marks G, W.K. Beatty. *Epidemics*. New York: Scribners, 1976
32. Ng D.K., C.H. Chan, R.S. Lee, L.C. Leung. "Non-contact infrared thermometry temperature measurement for screening fever in children." *Ann Trop Paediatr.* 2005;25(4):267-75. 13.
33. Ng E.Y.K, G.J. Kaw, W.M. Chang. "Analysis of IR thermal imager for mass blind fever screening." *Microvasc Res.* 2004;68(2):104-9.
34. Ng, E. Y.K. "Thermal Imager as Fever Identification Tool for Infectious Diseases Outbreak." *Medical Infrared Imaging Principles and Practices*. Ed. Mary Diakides, Joseph D. Bronzino, and Donald R. Peterson. Boca Raton,FL: CRC, 2013. 24-1-4-19. Print.
35. Norton, Paul R., Stuart B. Horn, Joseph G. Pellegrino and Philip Percoti. "Infrared detectors and detector arrays." *Medical Infrared Imaging Principles and Practices*. Ed. Mary Diakides, Joseph D. Bronzino, and Donald R. Peterson. Boca Raton,FL: CRC, 2013. 24-1-4-19. Print.
36. "No Touch + Forehead Thermometer - NTF3000." Braun, n.d. Web.
37. Powers, D M.W. "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION ." *Journal of Machine Learning Technologies*, vol. 2, no. 1, 2011, pp. 37–63., www.bioinfo.in/contents.php?id=51.
38. Pušnik I. and A. Miklavec. "Dilemmas in Measurement of Human Body Temperature." *Instrumentation Science and Technology*, (2009) 37:5, 516-530, DOI: 10.1080/10739140903149061
39. Ring, E. F. J., and K. Ammer. "Infrared thermal imaging in medicine." *Physiological measurement* 33.3 (2012): R33.
40. Ring E. F. J., A. Jung, B. Kalicki, J. Zuber, A. Rustecka and R. Vardasca. "Infrared thermal imaging for fever detection in children" *Medical Infrared Imaging* 2nd edn (Boca Raton, FL: CRC Press) (at press)
41. Rodriguez, J. D., A. Perez and J. A. Lozano. "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 569-575, March 2010. doi: 10.1109/TPAMI.2009.187
42. Ronneberger, Olaf, Philipp Fischer, Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, 2015, pp. 234–241., doi:10.1007/978-3-319-24574-4_28.
43. Rosenau MJ, Last JM. Maxcy-Rosenau "preventative medicine and public health." New York: *Appleton-Century-Crofts*; 1980
44. Roth, Joachim. "Fever: Mediators and Mechanisms." *Inflammation - From Molecular and Cellular Mechanisms to the Clinic*, 2017, pp. 861–890., doi:10.1002/9783527692156.ch33.
45. Rumelhart, D. E., Hinton, G. E., and Williams, R. J. "Learning representations by back-propagating errors." *Nature*, (1986) 323, 533--536.
46. Shlens, Jonathon. "A Tutorial on Principal Component Analysis." 25 Mar. 2003, www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf.

47. Simonyan, Karen, and Andrew Zisserman. "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION." *ICLR*, 2015, arxiv.org/pdf/1409.1556v6.pdf.
48. Smale, Alan. "The Electromagnetic Spectrum." NASA, Mar. 2013, imagine.gsfc.nasa.gov/science/toolbox/emspectrum1.html.
49. Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." *J. Machine Learning Res.* 15, 1929–1958 (2014).
50. Sun, G., T. Saga, T. Shimizu, Y Hakozaiki and T. Matsui. "Fever Screening of Seasonal Influenza Patients Using a Cost-Effective Thermopile Array with Small Pixels for Close-Range Thermometry." *International Journal of Infectious Diseases*, Elsevier, 20 May 2014, www.sciencedirect.com/science/article/pii/S1201971214014957.
51. Szegedy, Christian, W. Liu, Y Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V Vanhoucke and A. Rabinovich. "Going Deeper with Convolutions." *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, doi:10.1109/cvpr.2015.7298594.
52. Taubenberger, Jeffery K., and David M. Morens. "1918 Influenza: the Mother of All Pandemics." *Emerging Infectious Diseases*, vol. 12, no. 1, 2006, pp. 15–22., doi:10.3201/eid1209.050979.
53. "Temporal Artery Temperature Detector". Francesco Pompei, assignee. Patent 6292685. 11 Sept. 1998. Print.
54. "The 2009 H1N1 Pandemic: Summary Highlights, April 2009-April 2010." *Centers for Disease Control and Prevention*, www.cdc.gov/h1n1flu/cdcreponse.htm.
55. Weckmann, S. "Dynamic Electrothermal Model of a Sputtered Thermopile Thermal Radiation Detector for Earth Radiation Budget Applications" Master's Thesis, Virginia Polytechnic Institute and State University, (1997) Blacksburg, Virginia
56. Widmaier, Eric P., Hershel Raff, Kevin T. Strang, and Arthur J. Vander. "Vander's human physiology: the mechanisms of body function." 2016. Boston: McGraw-Hill Higher Education.
57. Wunderlich C. "On the temperature in disease; a manual of medical thermometry." *Univ. Leipzig*. Translated by W Bathurst Woodman, New Sydenham Society, London 1871.
58. Yao, Yi, and Gianfranco Doretto. "Boosting for Transfer Learning with Multiple Sources." *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, doi:10.1109/cvpr.2010.5539857.
59. Yildizyan, Aleksan, Jiawei Hu, Charles Squires, and James Gorsich. "Non-contact Medical Thermometer with Distance Sensing and Compensation." *Kaz Usa, Inc*, assignee. Patent 20140140368A1. Apr.-May 2014. Print.
60. Yildizyan, Aleksan, and James Gorsich. "Medical Thermometer Having an Improved Optics System." *Kaz Usa, Inc*, assignee. Patent 20140140370. Apr.-May 2014. Print.
61. Zhou B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. "Learning Deep Features for Discriminative Localization." *CVPR*, 2016 (arXiv:1512.04150, 2015).
62. "2009 World Population Data Sheet." *Population Reference Bureau*, Population Reference Bureau, 12 Aug. 2009, www.prb.org/Publications/Datasheets/2009/2009wpds.aspx.

B. Glossary of Terms

Acclimation time – time it takes for the system to reach equilibrium in a given ambient environment. For this research the primary focus is on physiological acclimation to the ambient temperature of their environment.

Ambient temperature – environmental temperature that the individual is in at the time of the measurement

Architecture – see Neural Network Architecture

Augmented data – increasing the inputted data by different functions (i.e. shifting the feature location in the image, flipping along the x and y axis, rotating the image 90 degrees)

Classification model – characterizing a data input into specific discrete categories (i.e. if the patient is febrile or afebrile from a temperature reading)

Filter – various sized matrix of values that get dragged across the output of the previous images, and produce the output to the next architectural layer of the neural network. They are the operator that provides the convolution in the convolutional layers of the network architecture. In conventional image processing they are known as **kernels**.

Hyperparameter – a value that provides input to the training of the deep learning architecture. There are a variety of different hyperparameters used for each learning architecture such as: learning rate, dropout regularization values, L1 or L2 regulation values, etc.

Image – Base data structure of a given sample set from a device (before any augmentation, windowing or external manipulation)

Input – input to the algorithm after external manipulation (i.e. after augmentation process)

IRFPA – Infrared Focal Point Array, an array of thermopile sensors that is capable of capturing a thermal data from the infrared spectrum emitted by a substance

Kernel – See Filter

Neural Network Architecture – The framework of layers (whether convolutional or fully connected layers) that is design to accomplish a given task, absent the training information. This is also referred to as the Architecture.

Neural Network – A neural network architecture that also contains the training information for the design

Noise – disturbance in the data. This can be artificially added to the input data or applied when lack of control of variables when gathering of the dataset. The latter was the objective to overcome for this research, and the uncontrolled variables were: thermal acclimation time, emotional status and facial orientation (the subject data later had a requirement of having both eyes present in the images used).

Regression model – characterizing a data input with a continuous value of base units (i.e. converting a patient reading to 98.6°F from a temperature reading)

Segment – extract the area of interest in a data input (i.e. extract the area affiliated with the eyes in an image of an individual)

Pooling layer – layer of a convolutional neural network where a mathematical operation decreases the resolution of a given layer. The common operations for this is max pool, or taking the maximum value of inputs, and average pool, taking the average value for a given area of inputs.

Pretrained model – Neural Network that has been trained on data from another dataset. This allows for the filters (that are the tools that allow objects to be identified), to be updated rather than generated from scratch.

Supervised Learning – Machine Learning techniques where the desired outputs are provided to the architecture so that the weights and biases of the layers (and filters in convolutional neural networks) can be updated to provide the desired results.

Reference devices – the device used to measure the accuracy or establish the baseline for the unit being tested. In the case of these experiments, the reference device was the Welch Allyn Suretemp 690plus.

Regulation – combating the network's tendency to be biased to the inputs of the training phase of supervised learning. This is an algorithmic decrease in dependency to allow the training loss to match the test loss more closely. Methods include dropout layers and L2 weighting for this research.

Tensor – multidimensional matrix

Vote – Algorithmic technique, using multiple inputs from the same image to give the primary decision for the image's febrile status. I.e. if the majority of the images were found to have febrile classifications then the final decision for that image would be febrile (and vice versa).