ASSESSMENT OF PROBABILITY CONDITIONS IN BINARY CLASSIFICATION SYSTEMS TO INCORPORATE AND LIMIT UNCERTAINTY IN OPTIMAL DECISION REGIONS

A Thesis

by

JOHN ANDREW BRINKLEY

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Douglas Allaire
Committee Members,	Raymundo Arroyave
	Richard Malak
Head of Department,	Bryan Rasmussen

August 2022

Major Subject: Mechanical Engineering

ABSTRACT

The study that is discussed in this thesis involves a unique method of quantifying uncertainty with respect to a classification problem. In essence, the objective involves redefining a materials classification problem pertaining to deleterious phases with respect to material composition and temperature as more of a function with inputs and outputs where the output is a probability label of either classification label that defines the probability of deleterious phases with respect to each of the aforementioned independent variables. This helps to interpret uncertainty in predictive statements that are assessed in a classification problem. The intention behind this method is to be able to set this type of system up as an optimization problem in order to maximize the likelihood of a desired condition, or minimize the likelihood of the undesired condition.

There are two primary approaches used in this study. One involves the use of a Gaussian Process Classifier to determine the aforementioned probability and discussing how to properly implement it and how to apply workarounds needed with the process. The other involves a more direct investigation of the data in what is called Sectioning and Proportioning, which involves taking the proportion of classification labels per section of the data to best assess the overall probability trend.

Both of these methods are found to have their strengths and weaknesses, and it is useful to use both in parallel with one another in order to assess any data that is being investigated while also interpreting it and adequately projecting the probability estimation as effectively and accurately as possible.

DEDICATION

To my mother Cindy, my father Paul, my sister Lindsey, my late paternal grandparents Bill and Sherry, my maternal grandparents Bob and Glenna. To all my Aunts, Uncles, and extended cousins who have been a part of my journey at Texas A&M.

ACKNOWLEDGMENTS

I would like to thank the Texas A&M University J. Mike Walker '66 Department of Mechanical Engineering for allowing me the opportunity to obtain a Master of Science Degree in Mechanical Engineering and extending to me a Graduate Assistant Teaching position for 3 semesters in a row. The experience and support from managing the courses I've served as a Teaching Assistant for has been invaluable.

I would like to thank Dr. Douglas Allaire of the Department of Mechanical Engineering for guiding and assisting me through my graduate research. His advice and direction through each stage of graduate school from the time before becoming my advisor to present has been extremely beneficial and has helped to guide my decisions at each juncture in graduate school. Additionally, the direction that had to be taken each week with respect to next steps and greater objectives of my research has been instrumental in the development of the research concepts that I have been investigating and working on.

I would like to thank Dr. Raymundo Arroyave and the other PIs of the Data-Enabled Discovery and Design of Energy Materials (D3EM) Program for extending to me the D3EM Fellowship opportunity in May 2020. The D3EM program not only specified the primary field of study I focused on throughout my time in graduate school, but also provided an environment for which my understanding and how much I was able to learn about the field of data science and materials science during that time has been absolutely invaluable as knowledge to carry forward.

I would like to thank Dr. Richard Malak and the other hosts of the weekly ESD meetings for providing the opportunity for collaboration with other graduate students about discussions regarding current hot topics in engineering as well as providing a platform that gave people the opportunity to both present their research as viable practice and as a way to gain good feedback as well as learn about the research and work being done by others.

I would like to thank my peers from graduate school who were both within my research group as well as others who worked in conjunction. Thank you for Danial Khatamsaz for you assistance

in helping me understand the concepts of Bayesian Optimization and Uncertainty Quantification, which has been able to help my research efforts significantly. Thank you to Richard Couperthwaite for the numerous pointers and assistance in helping me to understand one of the more predominant python modules known as George.py used for Gaussian Processes, which has greatly enhanced my ability to execute research objectives. Lastly, I want to thank Marshall Allen, whom I worked with as a team for the term project in my materials informatics course, and who's collaboration, experience, and resources helped me move my research forward in the most significant direction that established the largest foundation for which the main content within this thesis is based on.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor Douglas Allaire of the Department of Mechanical Engineering, Professor Raymundo Arroyave of the Department of Materials Science and Engineering, and Professor Richard Malak of the Department of Mechanical Engineering.

The data analyzed in Chapters 2 and 3 was provided by Marshall Allen of the Department of Mechanical Engineering, who obtained the data from Professor Richard Malak of the Department of Mechanical Engineering, which was obtained from a Thermo-Calc database.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by the Data-Enabled Discovery and Design of Energy Materials (D3EM) Fellowship sponsored by the National Science Foundation and by 3 consecutive Graduate Assistant Teaching Positions for the J. Mike Walker ’66 Department of Mechanical Engineering at Texas A&M University.

NOMENCLATURE

FGM	Functionally Graded Materials
DED	Directed Energy Deposition
GPC	Gaussian Process Classifier
BO	Bayesian Optimization
FGM	Functionally Graded Materials
$\mathcal{N}(\mu, \sigma^2)$	Normal Distribution
p(x|y)	Conditional Probability
$\propto$	Proportional to
GP	Gaussian Process
$\sim$	Similar to
$\kappa(x, x)$	Kernel Function
K(x,x)	Covariance Function
x_k	Subsection of X
C	Conditional label
A	Reliability factor
x_{km}	Midpoint of all x values in section interval
C_c	Number of Classifier Conditions
ϵ	error
S	Accuracy Score

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Concept of Uncertainty

The presence of uncertainty in almost any problem that can be conceived is an important factor to not only take into consideration but also understand the extent of. According to Chapter 2 of Uncertainty Quantification of Composite Structures[2], uncertainty can be divided into 3 broadly defined subcategories: inherent variability, lack of knowledge, and prejudicial uncertainties which consist predominantly of systematic and random errors. The first and last of these categories can generally be linked or grouped together more closely than either could be grouped with the second category. In other words, the second category is an important metric to both keep in consideration and, as often as possible, should be tracked throughout the process to be able to indicate regions within the design space of limited understanding. For the sake of simplicity, in this context the two primary kinds of uncertainties to be considered involve the lack of knowledge, and the lack of control. The former is fairly self-explanatory, the latter refers to how much inherent variability exists in the output under certain conditions, examples of this could include any measurement or human error as well as factors that lie outside of user control.

Arguably, the best way to model a system like this which includes inherent uncertainty in the output of the data is through a Gaussian process, in which instead of an explicit output that exists with respect to a set of independent variables expressed as a function y = f(x), instead there exists a normal distribution as the output as opposed to an explicit output y. This normal distribution exists in the form of Equation 1.1. The rationale behind this approach is to incorporate an inherent uncertainty in the system, where at each x location, there exists a range of possible values for the corresponding y output that could be modeled into a probability distribution with a mean and standard deviation, which varies with x. These factors, particularly the variance, can vary significantly with respect to the two aforementioned factors being lack of control and lack of knowledge.

$$p(f|X) = \mathcal{N}(f|\mu,\, K) \tag{1.1}$$

A condition where lack of control is present involves low correlation between x and y. In other words, if the range of any possible y values given the condition of specific x values is relatively large, then that generally indicates little correlation between the x independent variable and y dependent variable. In a practical sense, if adjusting a specific setting does not appear to have an effect on a desired output, then that is where a process would incur high variance, because there are a greater number of possibilities in that range. The ideal scenario would be to establish small ranges of possibilities, because that indicates a high level of control between the independent variable(s) and the dependent variable in question.

As for lack of knowledge, a new challenge is posed because not only is the level of control that exists between the independent and dependent variables not established, but any sense of a distribution is unknown at those conditions. Therefore, the prediction of a normally distributed range of possibilities under those conditions has to incorporate the lack of knowledge that exists there, which is compounded on the lack of control. The factors that can help achieve this incorporate assumptions based on existing knowledge. One of the most common approaches to finding this balance is through the use of Bayesian Optimization, an example of which using a simple mathematical equation can be observed in Fig. 1.1.

Bayesian Optimization in this context is primarily focused around model improvement. According to Chapter 1.1 of A Tutorial on Bayesian Optimization of Expensive Cost Functions[3], the premise of modeling the aforementioned probability distributions is by assessing the output as a posterior probability. The way that is created is by setting it proportional to two particular quantities, prior probability and likelihood, in a setup like what is shown in Equation 1.2, where M refers to the model and E refers to the evidence.

$$P(M|E) \propto P(E|M)P(M) \tag{1.2}$$

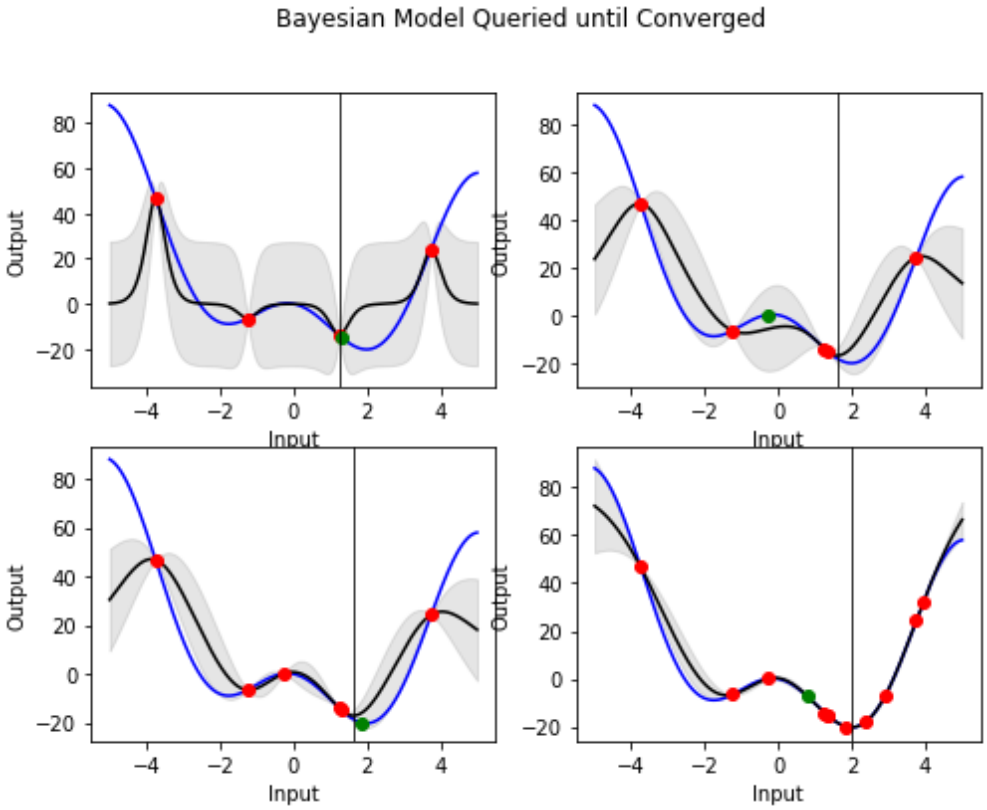

Figure 1.1: 2D Example of a Bayesian Optimization using an example ground truth function with new data points queried at each iteration until the model was assessed to have met optimal improvement conditions.

The model, while identifying regions of uncertainty, also incorporates a factor known as an acquisition function, which is projected as a function of x to identify which value will result in maximum model improvement if queried and implemented as a more certain data point in the next iteration. This acquisition function is defined as the probability of improvement according to Chapter 2.3 of A Tutorial on Bayesian Optimization of Expensive Cost Functions[3]. This process is defined in Equation 1.3.

$$PI(x) = P(f(x) \geq f(x^{+})) \tag{1.3}$$

In the context of this particular study, however, rather than optimizing the model through the iterative use of Bayesian optimization, the objective primarily involves creating a surrogate model sufficient enough to represent the trend expected by the data given the information available. Because of this, the acquisition function and the concept of querying new points are not as relevant in this particular study, at least in its current stage.

1.2 Gaussian Processes

There are a number of different approaches that can be used to find the desired optimal decision region from mentioned previously. All of the ones used in this particular study, to one extent or another, use a Gaussian process. To provide some insight in how this process works, it is essentially a generalized Gaussian probability distribution represented as a relatively ambiguously defined function of the independent variables[1].

Imagine a simple one-dimensional function that has a single input and single output modeled as a function y = f(x), where f(x) is not known and may not even be explicitly defined. This is practical in situations where either the independent variable, x, has limited influence over the dependent variable, y, and can only be used to reasonably estimate a range of possible values instead of giving one explicit value. The other scenario where this is useful is in a Bayesian Optimization process where there is unknown regions of the function space between x and y.

The application of a Gaussian process here is to create a representative, or surrogate, model

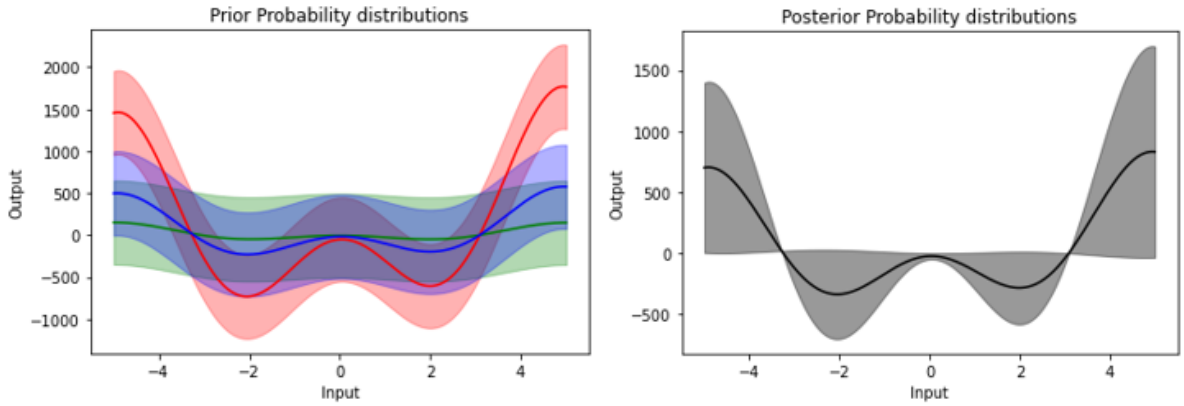

Figure 1.2: left: three samples from the prior probability; right: two datapoints are observed, a mean prediction trendline with a shaded variance range containing the 3 prior samples projected as a posterior probability[1]

.

of what the dependent variable is expected to be as a function of the independent variable. This model is a probability distribution indicating the range of possible y values at any x location. This distribution is, in this context and in general, normally distributed with an estimated mean and standard deviation, the value for each exists as a function of x. These values are determined through a process of assessing the prior and posterior probabilities, a depiction of which is displayed in Figure 1.2.

In order to model this as a Gaussian process, the intention is to represent this as a probability function with a mean and standard deviation. In the format of a function, this is quantified in the form of a mean function and covariance function[1]. The function established is defined in Equation 1.4.

$$f(x) \sim GP(m(x), k(x, x')) \tag{1.4}$$

The covariance function acts as a matrix when incorporating multiple different basis functions in order to establish a surrogate model that emulates Fig. 1.2. In practice, a kernel function is used in order to establish the covariance between functions. According to Chapter 6 of Pattern

Recognition and Machine Learning[4], there are a number of different functions that can be used for this kernel function including but not limited to Radial Basis Function, Squared-Exponential, Matern, Linear Regression, Automatic Relevance, and Nadaraya-Watson Model. These kernel functions are incorporated into the gaussian process using Equation 1.5.

$$\begin{bmatrix} f \\ f_* \end{bmatrix} = \mathcal{N}\left(0, \begin{pmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{pmatrix}\right) \tag{1.5}$$

1.3 Materials Background

The data used in the study of further understanding and incorporating uncertainty into a system design is through the use of computationally graded alloys, which act as a subclass of functionally graded materials, or FGMs[5]. These materials are created using a Directed Energy Deposit, or DED, process which presents the ability to easily change material composition layer by layer by depositing powders of user specified compositions and solidified using a high energy laser[5]. However, one of the obstacles to this approach is that periodically, material phases can become present in the micro-structure that lead to undesirable properties or cracking. These kinds of phases are referred to as deleterious phases, and should be avoided in design from these kinds of materials.

The objective of this study is to avoid deleterious phases. To establish how they can be avoided, first the variables within user control need to be identified. As an example study, one particular set of data contains phase composition information of 4 different elements: Iron, Nickel, Chromium, and Titanium. Those are listed along with a corresponding temperature measurement that ranges from 300K-1800K. This dataset is generated externally through the use of Thermo-Calc software [6]. The ideal setup of this would be to have multiple forms of test data of materials for this study. However, considering the relatively statistical scope of this study as well as the extent of the data necessary for an adequate analysis, the simulation will suffice. The data outputted through this process involves the 5 aforementioned variables, being the 4 material compositions plus temperature, and a classifier label of 1 or 0, 1 indicating the presence of a deleterious phase, and 0 indicating a lack thereof. The objective is to find an optimal decision region that both

minimizes the uncertainty of the assumption that a deleterious phase will be present at a given location, while also large enough such that the variable space available to optimize for any other desired properties in another study is maximized.

1.3.1 Similar Studies

It is viable to observe how Gaussian processes are incorporated in previous studies as well as in which manner they are utilized. In the article Gaussian Process Surrogates for Modeling Uncertainties in a Use Case of Forging Superalloys[7], a Gaussian process is implemented to create a surrogate model to act as a potential replacement for expensive FEM simulation. The rationale was to accommodate for both the time and computational expense that is required by FEM, and bypass that through the use of a Gaussian model to acquire a surrogate model fit for the expectation at each point, which varies with less knowledge at each point. This is a viable study for assessing the strength and effectiveness of the Gaussian process especially with existing FEM data and the ability to acquire direct testing results to compare the results of the process with. This, however, presents an advantage that is not present without a ground truth model since various parameters have to be modified to fit the expectation accordingly.

According to Gaussian Process Kernel Transfer Enabled Method for Electric Machines Intelligent Faults Detection With Limited Samples[8], the most effective kernel function with the highest detection accuracy was the Radial Basis Function, or RBF, kernel. This had at least 5% greater accuracy used than other kernel functions. Therefore, as a kernel study it is reasonable to assume that the RBF function will suffice as arguably the best kernel function for general use practices including the application analyzed in this study. However, it is useful to implement others as well in other applications to see if they perform better in some areas more than others.

The article titled Efficient Global Optimization of Expensive Black-Box Functions[9] discusses the use of stochastic processes to assess the probability distributions of different possible outcomes and emphasizes the need to perform cross-validation in order to verify the models. The probability distributions in this particular study are more unique to quantify because they come in two forms: the probability of each classification case, and the probability distribution of each corresponding

probability measurement across each point. The latter of the two is more relevant when measuring proportions from the data directly. Cross-validation is important and used extensively in this study, particularly because of the limitation that exist in the computational ability of the software modules. The rationale for cross-validation becomes more important with a limitation of the size of the training data that is able to be used because it encapsulates a limited scope of the full data. Therefore, this means multiple scopes have to be run and compared in order to obtain a reasonable estimate.

2. DATA SECTIONING AND PROPORTIONAL MEASUREMENT

2.1 Objective

The objective of this study is to create a probability metric that acts as a function with respect to the independent variables in a classification problem. This probability metric represents the chance that one of the two classifications could be true in a binary classification problem. This is a useful quantification when there is uncertainty present in the data because a simple step function that assesses whether certain combinations of variables correspond to one binary condition or another cannot always be conclusive and the possibility is there that the label is inaccurate for a number of reasons.

One simple example could be a question of basketball skill, for instance. If one wanted to assess how the independent variable of height corresponded to the probability that the person will make a majority of baskets they attempt, what kind of trend could be established? If one ran a test where they took one person 5 feet tall and one person 6 feet tall, and the 6 foot person makes a majority of baskets out of 10 shots taken while the 5 foot person does not, is that conclusive data that greater player height will correspond to the condition that a majority of shots taken will be made successfully? Based on this experiment alone there is nowhere near enough of a test to draw that conclusion, mainly because the sample size is so small and there are so many possible factors in play that are not considered in this test that easily correspond to sources of uncertainty. Hand eye coordination, weight, and overall athleticism are just some of the factors not quantified in this study that could very well play a factor in the outcome. Some of these factors are more difficult to quantify than others, especially ones with broad definitions such as overall athleticism.

This may seem like an oversimplification and obvious statement. However, the factors in play in that oversimplification are found to absolutely be present even with large amounts of knowledge, albeit on a much smaller scale. There are two main factors that are primarily taken into consideration here; the first is the reliability of the information present from the given samples

available, the second is the availability of information present from the given samples. These are important because they correlate to how uncertain the data is. If the given information is not absolutely conclusive, then it has to be inferred then there is a degree of uncertainty present in the data. Quantifying this is easier in some cases than in others. If the output is a specific value, such as material yield strength, then that is relatively easy to quantify because one could run a given number of tests on samples with the same independent variables under analysis and then model the results as a normal distribution where the mean and standard deviation could be obtained under those given conditions. In other words, it would be relatively easy to model that system into a Gaussian process. However, when it is a classification problem where the output is a binary classification label, such as a case where an assessment is made on whether or not a system works, then the uncertainty in that prediction is more difficult to quantify and assess and can require some assumptions to be made.

2.2 Sectioning and Proportioning

This method of assessing the probability as a metric variable in this analysis involves a more direct approach of sectioning the data and proportioning the samples within. This is a relatively simplistic approach but for the objective of propagating the probability metric out into the data to directly find variable regions that correspond to the desirable conditions that the user wants to obtain, it is useful and also provides some information that cannot be inferred through using a more thorough albeit automated process like a Gaussian process classifier. To assess this method, the dataset used contains a classifier label indicating the presence or lack thereof of deleterious phases corresponding to 4 material compositions: Iron, Nickel, Chromium, Titanium; and an ambient temperature value. This data contained 50,000 total data points and is produced by a simulation run from Thermo-Calc software [6].

To start, imagine taking the full proportion of deleterious phases over the entire dataset. This creates a broad probability metric where the condition is "If a random phase is taken, then there will be a value P representing the probability that it will be a deleterious phase". The condition assessed in that probability metric is less important, so one could also create a metric for the opposite

condition. The only convention to keep in mind is whether or not the identified probability metric should be minimized or maximized. In this case, since deleterious phases are to be avoided, the probability of a deleterious phase being true under given conditions should be found at its minimal values. That said, now that the convention for identifying the probability metric as a proportion of the data; which is assumed to include some inherent uncertainty, more on that later; the metric can be split between sections of the data in order to create conditional probability metrics. These sections of the data appear as subsections of the full dataset as shown in Equation 2.1.

$$\mathbf{x}_k = \{X_i\}_{i=k}^{s(k+1)} \tag{2.1}$$

For simplicity purposes, by beginning with one independent variable, Iron in this case for demonstration purposes, by sectioning iron in two groups, the probability of a deleterious phase within each interval can be estimated by finding the proportion of deleterious phases that exist within. The iron variable in this case is sectioned by a composition value within the bounds of [0,0.5], and [0.5,1]. From there, the proportion of deleterious phases can be assessed within each condition, where metrics are created that state that where Iron composition is between 0 and 0.5, the probability of a deleterious phase is P1, and where Iron composition is between 0.5 and 1, the probability of a deleterious phase is P2. This concept allows for the metric to be propagated throughout the dataset.

The same practice can be done by sectioning Iron into 5 intervals, 10 intervals, 20, 50, 100, and so on to further propagate the metric. As the number of intervals increases, the ranges of Iron values in each interval decreases by scale. This can be shown in Fig. 2.1. For each vector of x_k as defined in Equation 2.1, there exists an equal length classifier label vector denoted as y_k, from which the proportion of deleterious phases can be found using Equation 2.2 where n is the number of samples per interval and y_i is one sample of y_k. The way that the classifier is defined is such that y_i = 1 when deleterious and 0 when non-deleterious.

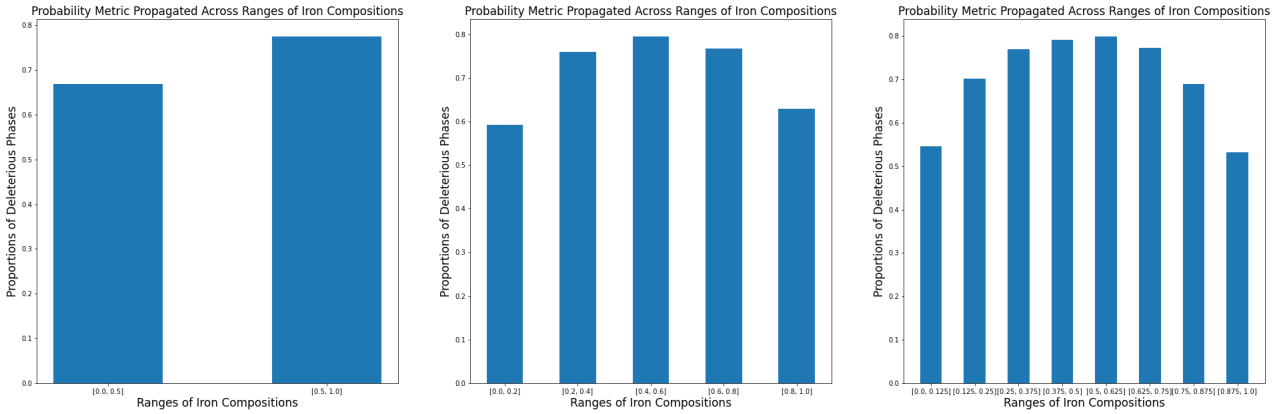

Figure 2.1: Bar Graphs Displaying the Proportion of Deleterious Phases as a function of Iron composition ranges with varying intervals. Left: 2 intervals, Middle: 5 intervals, Right: 8 intervals.

$$P = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \tag{2.2}$$

2.2.1 Multi-dimensional Sectioning

Now that the sectioning process has been introduced, it is time to expand the number of dimensions used in this process. When multiple independent variables are considered in a study like this, there are multiple sections that can be made. Starting with the one-dimensional sectioning process, one could limit a variable such as Iron composition to be within a set of bounds in any given interval, but all other variables still have full range of possible values that fit within the established constraints. In other words, there still remains a high degree of ambiguity when it comes to the number of external factors that could be influencing any variables used in this study. Therefore, to reduce that ambiguity the sectioned are expanded into a greater dimensional space.

Starting with a two-dimensional approach, in this case Iron and Nickel, in each interval of Iron that was established in the one-dimensional approach, the same number of intervals for the second independent variable, Nickel, can be taken and the proportional measurement of deleterious phases within each found and denoted with respect to each interval. This creates a sort of two-dimensional step function where the proportional measurement, used to estimate the probability of a deleterious

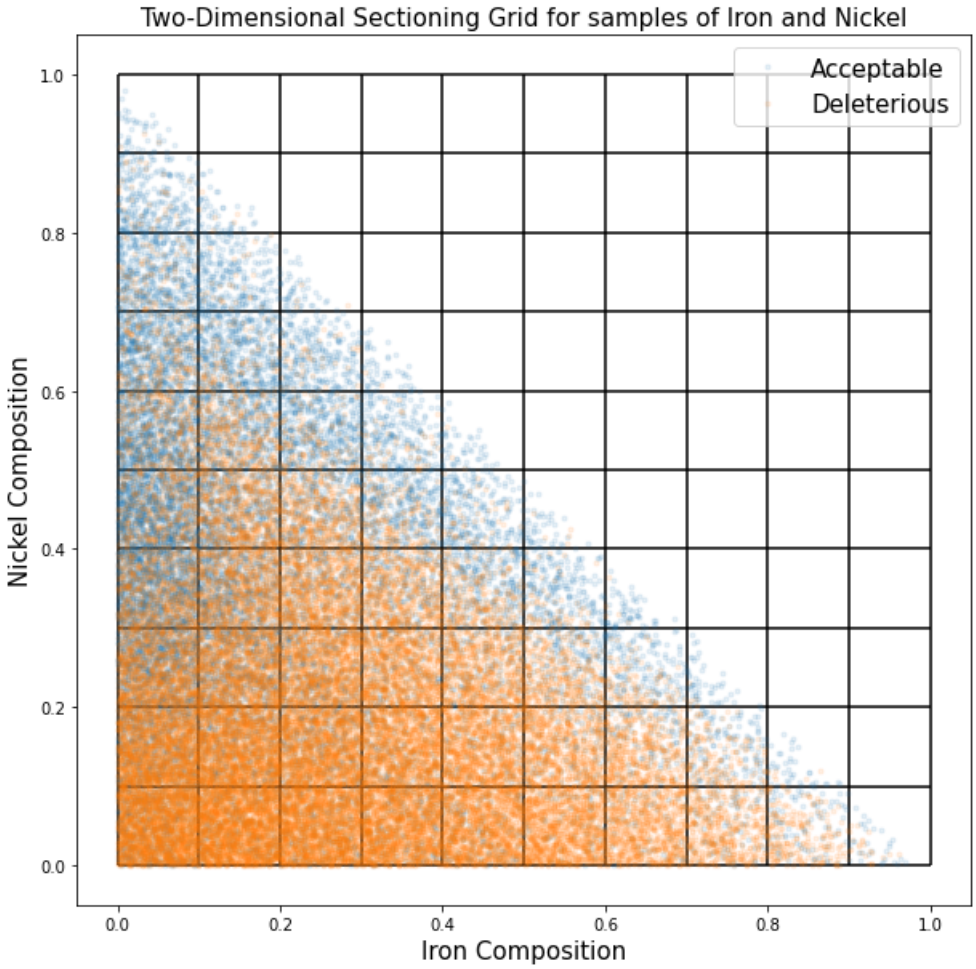

Figure 2.2: Two-Dimensional Projection of Compositions Iron and Chromium along with their corresponding acceptable(blue) or deleterious(orange) classifier label shown with a grid indicating 100 total intervals of variable ranges within which the proportional measurements of deleterious phases can be found.

phase being present, is the output from each interval. This process is illustrated in Fig. 2.2. This process can be expanded into 3, 4, and even 5 dimensions. However, it is important to note that many of the intervals in this projection are empty. The reason for those intervals being empty is because they exist within a space that violates the constraint boundary of the variables used in this study, which is an important factor to keep in mind.

2.2.2 Constraint Boundary

Given that the variables of Iron and Nickel as well as Chromium and Titanium represent material compositions, a constraint exists in this study where each of the 4 compositions have to sum to 1, it would be impossible not to. When sectioning the variable space, this can prove to be a challenge, especially when it comes to assessing uncertainty with the proportional measurements taken from each. In the prior two-dimensional example it is relatively easy to assess this given that it exists within a linear boundary that bisects the variable space into 2 equal sized regions. In this case is a simple line of y = -x, in which samples cannot exist to the right or above that line. The main consideration is how this constraint boundary intersects any of the intervals used in this study. Fortunately, in this two-dimensional case it only affects the intervals that exist along the line y = -x as displayed in Fig. 2.2, within each there is only a 50% intersection of the constraint boundary. This comes into greater consideration when assessing the reliability of the proportional measurement within that specific interval, more detail on that later.

As the dimensionality of this sectioning process increases, however, the type of intersection that the constraint boundary will have with the interval will become much more difficult to determine, especially when the dimensionality exceeds 3 dimensions, in which case the visibility of the process becomes impractical. Unlike the two-dimensional example, it wouldn't be as simple of a case of where each interval that sits on the boundary line has an equal and easy to see fraction of intersection with the constraint boundary. In 3 and 4 dimensions it can and is found to vary from interval to interval. Therefore, when it comes to assessing this fraction of intersection within each interval, it is practical to have a general convention for how to accomplish this. One relatively simplistic method that uses this same sectioning concept can be useful for this.

The way that this sectioning approach would work is that within each interval section established previously, first and foremost each one has to be assessed to ensure that samples exist within, otherwise it is neglected entirely. After that, within each section the user can create multiple subsections using the same approach as before on a smaller scale. However, the only difference in this case is that instead of assessing the proportion of deleterious phases within each subsection, the

constraint boundary is assessed in each subsection. The way this constraint boundary functions is very problem specific, and will change with changing dimensionality of the interval sections. In this case, since all 4 compositions have to sum to 1, a case where all 4 compositions are present will be assessed. The only factor sought after in this case is whether or not the subsection contains the constraint boundary. Each subsection is defined by a lower bound and an upper bound for all the independent variables in the test. When there are 4 variables, the condition is met under the circumstance shown in Equation 2.3, where C is the conditional label for each subsection, n is the number of variables used, x_L refers to the lower bound vector of each independent variable, and x_U refers to the upper bound. However, when less than 4 variables are used, or when less than all 4 compositions in this study are used, then the boundary condition is slightly different where instead of ensuring fit, the objective is to only look for conditions where both bounds are greater than 1, because when they are less than there still exists an unbounded third and/or fourth variable that can reach that value. A 2D representation of how this works can be seen in Fig 2.3.

$$C = \begin{cases} 1, & \text{if } \sum \mathbf{x}_L < 1, \sum \mathbf{x}_U > 1, n = 4 \\ 1, & \text{if } \sum \mathbf{x}_L < 1, n < 4 \\ 0, & \text{otherwise} \end{cases} \tag{2.3}$$

From here, the fraction of intersection between each interval and the constraint boundary can be assessed. Using Equation 2.3, one could generate a C conditional value for each of the subsections present in each interval. For example, if there were 100 subsections, then there would be 100 corresponding C values, each being equal to 1 or 0. From there, the intersection fraction can be assessed for each interval by taking the sum of all the C values and dividing it by the total number of subsections using Equation 2.4, where F is the intersection fraction, n is the number of subsections, i is the subsection index.

$$F = \frac{1}{n} \sum_{i=1}^{n} \mathbf{C}_i \tag{2.4}$$

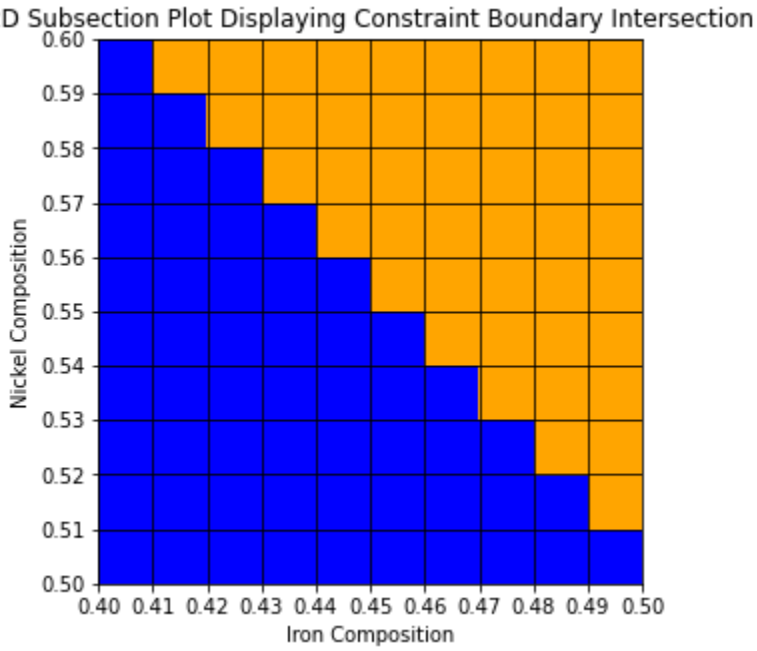

Figure 2.3: Two-Dimensional Subsection plot with 100 sub-intervals within the interval from 0.4 < Fe < 0.5 and 0.5 < Ni < 0.6. Blue indicates that the sub-interval satisfies the constraint boundary, whereas Orange indicates that the sub-interval does not.

This process is repeated for each of the original section intervals originally established. While this process presents a relatively rough estimation for the fraction of intersection between each interval and the constraint boundary, it can converge onto the true value as the number of subsections increases. Imagine the use of block shaped pixels to create an image, the smaller in size they are the greater detail that exists in the image. However, the extent to which this is created has to be taken into consideration with the capabilities of the software being used. If the four composition variables are taken into consideration with 10 intervals each, then that creates 10^4, or 10,000 intervals, with however many subsections in each interval that lies within the constraint boundary. Increasing the number of subsections will significantly increase the computation time to find the fraction of intersection with the constraint boundary for each. So if the need to be thorough is not as important, or a relatively rough estimation will suffice, then a relatively smaller number of subsections that don't drastically increase computation time beyond reason will be acceptable. This fraction of intersection becomes more relevant when determining the reliability of the probability

metric created from each interval, more on that later.

2.3 Minimum Effective Section

As was previously expressed, the size of the sections created in this process can be scaled from 2 large sections that bisect the data to 100 small sections that propagate the probability metric into smaller margins. However, based on the scope of the data available, there is a limitation to how small these margins can effectively go. This concept is similar to the practice of histogram smoothing [10] in that the ground truth model is assumed to be a smooth curve fit, and the data distribution used should represent that as best as possible while also propagating as much of the probability metric as possible. The difference here is that instead of using this concept for a histogram representation of a probability distribution, it is used to assess the probability metric estimated at each interval by the proportional measurement. This is where the limitation of how much the available data shows the user comes into effect the most.

The ideal scenario would be that at any specific combination of independent variables with common values, or at least in close enough proximity to one another, produces enough test result points such that an adequate probability estimation under those specific conditions can be assessed. However, the challenge here is that data is not readily available in that kind of proximity. Therefore, in order to assess a reasonable probability estimation under those conditions, the proximity region has to be widened enough such that it includes enough points to be able to assess an estimated probability metric from the proportion of deleterious phases within each interval.

Finding this minimum effective section could be an optimization problem in another study. The main objective in this study is to ensure that the sections of data used are as small as possible while still containing enough points to make an adequate prediction of the probability metric. If the intervals are small enough such that they only contain a single digit number of points, then a major limitation is presented on the study because there is very little margin for error in those conditions, and the possible proportion values that can come from a small number of data points is limited. For example, in an interval that only contains 2 data points, the only possible proportion values that can be found are either 1, 0.5, or 0.

Having so many of these limited range values in the greater variable space after the fact make it so that the metric is not sufficiently patterned out, and a reasonable estimation of the ground truth probability metric that is being sought after cannot be obtained when the intervals are this small. Not to mention, a higher number of intervals also means longer computation time to the point where solutions couldn't be obtained within a reasonable time frame. In other words, they could be made small enough such that they could take several hours or even a day to produce results, ones that would be unreliable in this case given the scope of the data available.

2.3.1 Correlation between Sample Count and Constraint Boundary Interface

For the purpose of this study, the convention is that approximately 50 points per interval is reasonable enough. The actual range of sample count between intervals varies quite significantly, a lot of which depends heavily on how much each interval intersects the constraint boundary. In other words, the intersection that exists between the constraint boundary and each interval makes up the space in which sample points are able to be present. The balance of these two factors allows for identifying regions of higher data density and lack thereof, which is hard to take into account directly in a classification approach. In order to take both of these factors into account, a direct proportionality is considered between the two factors because with a larger region of data point availability, the greater the number of data points could theoretically be present in that interval.

This is a direct one to one correlation, which is deemed acceptable because if you imagine bisecting a cube in two, then half of the space is available, within which half the amount of points out of the total number of points assuming uniform distribution could theoretically be present. By bisecting it again, then that goes from half the number of points to now a quarter of the total number of uniformly distributed points. Therefore, with that convention, a factor can be established that assesses how much data exists within a given point considering both the number of data points and the constraint boundary intersection. This factor can be displayed in Equation 2.5, where A is the sought after factor for any given interval, N is the number of sample points in the same corresponding interval along with F as the fraction of intersection with the constrant boundary previously established in Equation 2.4.

$$A = \frac{N}{F} \tag{2.5}$$

The actual numerical value of this factor is less important, as it is intended to be used as a scaling factor to assess the reliability of the proportion measurement as a probability metric estimate, more on that later.

2.4 Gaussian Process Regression

One important factor to keep in mind is that the proportion of deleterious phases per interval cannot be automatically assumed to be an exact representation of the probability of deleterious phases under the same conditions. In other words, there is some inherent uncertainty in those measurements. Quantifying that uncertainty, however, is a challenge because as a classification problem, it is difficult to assess the probability metric in a way that wouldn't be present if it was some arbitrary number. If it were some arbitrary number, then determining the uncertainty would be relatively easy because it can be assumed that at each point there exists a normal distribution of values for which a probability distribution with a mean and standard deviation could be established. That is not the case for an actual probability metric because it is derived from output data that is explicitly a classifier label of 1 or 0. Because of this, the definition of the uncertainty at each measurement is arbitrary and requires some assumption.

The use of a Gaussian process in this case can be useful at not just incorporating that uncertainty, but also create a metric that serves as more of a direct y = f(x) function of each independent variable rather than a step function as defined previously. To provide some background, a Gaussian process essentially represents a probability distrbution at each variable location showcasing a normal distribution of possible values that an output can be given the conditions established by the independent variables[1]. This normal distribution can be observed in Equation 2.6 for any given combination of independent variables x_k. In other words, f(x) is a surrogate model that represents a probability value.

$$p(f(x_k)) = \mathcal{N}(\mu_k,\, \sigma_k^2) \tag{2.6}$$

When it comes to creating the surrogate model using the Gaussian Process Regression approach, first and foremost the intention is to develop a new set of data points with the probability metric as opposed to the original set of data points with a classifier label since this is a regression approach. The original data points are incorporated into a similar study that will be analyzed in the subsequent chapter using a Gaussian Process Classifier. This study involves a more direct regression approach. Therefore, in order to accomplish that, it is important to have an explicit representation of the independent variable vector corresponding to each proportion value from each interval. This is unique to each individual section because since the interface between each section and the constraint boundary varies by section, more so with increasing dimensionality.

2.4.1 Center Point Representation

The best way to describe this independent variable vector that corresponds to each probability metric is essentially a center of mass of all of the data points within each interval. This is an important factor to quantify correctly because if a general assumption is made then the surrogate model created will be significantly off especially since a number of cases could have results that lie outside the constraint boundary. Each previously established interval is defined by a lower bound and upper bound or each independent variable. The next step in this process is to transition this from a step function per interval to more of a smooth curve function fit. Therefore, the first step in that is to have the proportion measurement set at the point that represents the center of mass in each interval.

The easiest and arguably most accurate way to determine the center of mass has to do with the samples present within each interval themselves rather than the available sample space because that can help accommodate non-uniform distributions of samples within each interval. Fortunately, this is as simple of a process as averaging each of the variables of all of the data points within each interval. This is deemed to be acceptable for two reasons: first, the result fits the constraint

boundary of each material composition summing to 1, and second, by averaging each of the data points present one can effectively arrive at a midpoint of the data within each interval. This process can be demonstrated by Equation 2.7.

$$x_{km} = \frac{\sum x_k}{n_k} \tag{2.7}$$

2.4.2 Reliability

As was mentioned previously, the reliability of the data within each interval is a factor that has to be taken into account. This is where that factoring in occurs. One of the most difficult factors to take into consideration is assessing the uncertainty in each of the probability metric predictions from the use of proportional measurements. As previously stated, this is not a system in which numerical outputs are used that can be easily formatted into a normal probability distribution at each point, this is essentially creating a normal distribution at each point using a probability metric itself. That fact alone means that certain conditions have to be satisfied, most importantly the fact that the probability metric has to be between 0 and 1, no exceptions. Additionally, it means that the uncertainty is very arbitrarily defined with this metric. Because of this, certain factors have to be taken into consideration.

The reliability of the data is one of those factors; this is previously quantified using the A value from Section 2.2.2. It is mentioned that the actual value of the factor is less important, the reason is because in this context it is used as a scaling factor for the error. The convention for this is very arbitrary and could be modeled in a number of different possible ways to see the effect that it has on the surrogate model. To recap from before, the reliability factor is a ratio between the number of data points in any given interval along with the fraction of intersection between that same corresponding interval and the constraint boundary. This allows for a proper assessment of how much available data exists in any given interval, and the reasonable assumption is that the more data available there is, the more accurate the proportional measurement will be when representing the probability metric in the data.

In this study, the scaling factor is directly applied to an arbitrary error estimation, which is user-defined and can be adjusted based on how it affects the surrogate model. In the convention used in this study, a linear relationship is used, which is essentially as simple of a function as ϵ = -ar + b, where r is the reliability factor, b is the intercept, and a is the coefficient which is multiplied to the reliability factor. These coefficients are ambiguously defined but they do have to be scaled accordingly. Since these errors are for probability estimations, they are limited in their scale and should generally expect a maximum value of 0.1, which can be scaled accordingly to observe the effects on the output. However, the minimum has to be no less than 0 no matter what, which makes this convention more important to established.

One other factor to consider is that the linear relation is based on assumption. In practice, it it worthwhile to consider other error scaling factors as well, including quadratic and exponential scaling. This falls into the category of future work to investigate in order to further establish this method to see if the convention is feasible for various applications, especially in comparison to a classifier approach. That said, one convention that should remain consistent is that the error should always decrease with increasing reliability of the data per interval because of the presumption that the more data available, the more confidently the probability under those conditions can be assessed based on the proportion, the less potential error there would be.

2.4.3 Temperature Dependency

The fifth variable under this study, which has not been focused on as much in this study up to this point, is the temperature value. This is because the effect that temperature has on the output is relatively unique in comparison to that of the four compositions, and this can be observed directly from the proportion measurements directly. As a variable that differs from composition and one that is not bounded by an explicit constraint boundary, it is difficult to incorporate this into the study interchangeably with any of the material compositions. Additionally, from initial classifier testing, incorporating temperature was found to cause the classifier to produce a poor fit, more on that in Section 3. However, because of that, temperature was not incorporated into this study initially. In this process, using the same interval procedure as before, temperature is easy to incorporate as a

fifth variable in the study. That said, it is important to take the results of the original 4 variable sections into consideration, particularly the proportional measurements.

Arguably the most important observation made from this practice, which carries directly into the observations from the subsequent classification approach, is the fact that within a substantial number of the material composition based intervals, there exists a proportional measurement of deleterious phases equal to 100%. This means that the variable temperature has no effect on the output probability in this case. However, in other cases, temperature does have an effect. This is a unique application because, in essence, temperature only acts as a step function where its effect on the probability metric will only be present under certain conditions. The next step involves assessing those conditions, which takes place in the subsequent classification chapter.

3. GAUSSIAN PROCESS CLASSIFICATION

3.1 Background information

The approach used here is arguably more efficient and thorough than the prior sectioning and proportioning approach from Chapter 2 in that instead of going through the step by step process of deriving a probability metric and inputting that into a Gaussian process regressor, instead this jumps right into using the Gaussian process from the start. This process involves the use of a Gaussian process classifier to determine the aforementioned probability metric and use it as a way to separate the data based on variable regions to identify an optimal decision region for both binary classification conditions. Using the same data as before, the Thermo-Calc deleterious phase labeled data under the 4 material compositions and temperature vector [6], these two conditions are classified as acceptable and deleterious. Therefore, the objective of this study is to find a region within which the probability of a deleterious phase is reasonably low enough for design purposes but also as large as possible to maximize the possible space that can be used for other forms of optimization.

The Gaussian process in this case uses a portion of Bayes theory where the output is the posterior probability which is derived based on the prior probability and a likelihood function. In a regular design problem that uses a Gaussian process, the prior probability can be assessed as a normal distribution present based on the data available whereas the likelihood function is a quantification on how likely the data at any given point is. The process of determining the posterior probability can be indicated in Equation 3.1 where M refers to a model and E refers to the evidence, which therefore allows for the conditional probability of the evidence given the model to be joined with the probability of the model to determine the posterior probability[3].

$$P(M|E) \propto P(E|M)P(M) \tag{3.1}$$

In a classification problem, this process is slightly redefined to fit a probability metric, in which

it incorporates the probabilities of the classification conditions. In a binary classification problem this is relatively easy because there are only two possible conditions. In essence, this helps to scale the posterior probability such that it represents a probability metric for one or both classification conditions depending on how the process is set. This can be observed in Equation 3.2, where C refers to the total number of classifier conditions and c refers to the classifier index[1].

$$p(y|x) = \frac{p(y)p(x|y)}{\sum_{c=1}^{C} p(C_c)p(x|C_c)} \tag{3.2}$$

3.2 Implementation

In this instance, a Gaussian process classifier module from scikit-learn is used in order to generate a probability metric and decision region output from the data. In order to configure this algorithm, a factor known as the kernel has to be inputted in order to set up the covariance matrix for the Gaussian process[11]. In setting up the covariance matrix, kappa is used as a parametric kernel function as represented in Equation 3.3.

$$K = \kappa(X, X) \tag{3.3}$$

$$\kappa(x_i, x_j) = \exp(-\frac{d(x_i, x_j)^2}{2l^2}) \tag{3.4}$$

The Radial Basis Function, or RBF, is used in this instance as the kernel function along with a starting multiplier that is user defined and can be changed. This function can be shown in Equation 3.4[11]. The benefit of this process is that the kernel will converge onto an optimal value through an iterative process that is automated. The only factor that needs to be user-defined is the multiplier. In order to create a condition that will accommodate this requirement, a counter needs to be in place that finds a condition under which if the fit is not ideal or the kernel is unable to converge properly, then it can detect that condition and set a new starting value for the kernel multiplier, there can be a couple of counters in place to detect this occurrence and try a couple of

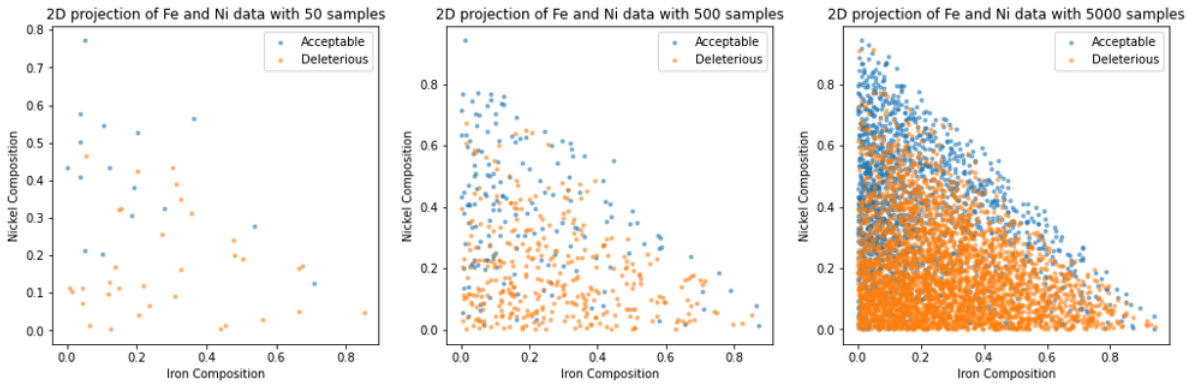

Figure 3.1: 2D Plots of Varying Sample Counts projected over Compositions Iron and Nickel. Left: 50 Data Points, Middle: 500 Data Points, Right: 5000 Data Points.

different multipliers to obtain more desirable results. This can be entirely user defined.

The crucial part of this study is inputting data into the Gaussian Process Classifier. The data used in this study contains 50,000 points, which is extensive enough such that, given the disposable tools, will override the classifier given its shear number and make it such that no output is able to be computed and instead the module will produce an error. Additionally, one of the additional risks of this is that with a large number of data points, the module will produce an output that assumes a more random distribution and less patterning because of the greater volume of data points that overlap with one another. In other words, in this type of classification problem the presumption is that both conditions have an inherent degree of uncertainty in the prediction, which means that there will be overlap between the two optimal decision regions and an error region. One of the challenges in this, however, is that with increasing numbers of data points, the amount of points that exist in those overlap regions increase substantially. An example of this occurrence can be shown in Fig. 3.1, where as the number of data points increase, the amount of overlap between decision regions increases as well. Therefore, in order to accommodate this limitation, a workaround has to be put in place that balances both the overall pattern while limiting the number of points in overlap such that the classifier can still produce an output representative of the full pattern while not overloaded or over-fit.

3.2.1 Training the Classifier

In order to properly train the classifier with this limitation, the number of training points used needs to be selected with both the balance of displaying a pattern representative of the full dataset while also small enough such that it doesn't override the classification algorithm or present enough overlap such that a random assortment is assumed. This exact number could be used as an optimization factor in another study. However, for the sake of this study the intention is to find something that works. The training points inputted into the classifier must meet the condition of $x_k \subset$ X, where X is the full dataset. Through trial and error, a value of 500 was found to suffice when used as a number of training data points for the scikit-learn Gaussian Process Classifier Module[11]. However, when using a number of 500 to create a surrogate model that is supposed to fit a pattern in which 50,000 total data points are present, not to mention to assess a reasonable predictive condition for those regions, it could be argued that the extent of the training data is not enough at this scale. With the technological limitation imposed as well, this merits the needs for a workaround.

In order to create this workaround, first the outputs of the Gaussian Process Classifier have to be assessed. There are two of interest in this case, one is the probability metric for each classification condition generated by the Gaussian process classifier, the other is the accuracy score. This accuracy score is a way of assessing how well the classifier fits the data, or in other words it can be expressed as shown in Equation 3.5, where ϵ is the error and S is the Accuracy Score. This quantity is one that, should be maximized as much as reasonably possible. The layout of this workaround can be shown in Fig. 3.2.

$$S = 1 - \epsilon \tag{3.5}$$

The theory being tested in the workaround in order to both accommodate the technological limitation while also including enough training data to adequately represent the pattern is to run multiple Gaussian process classifiers on independent random sections of the data with equal length,

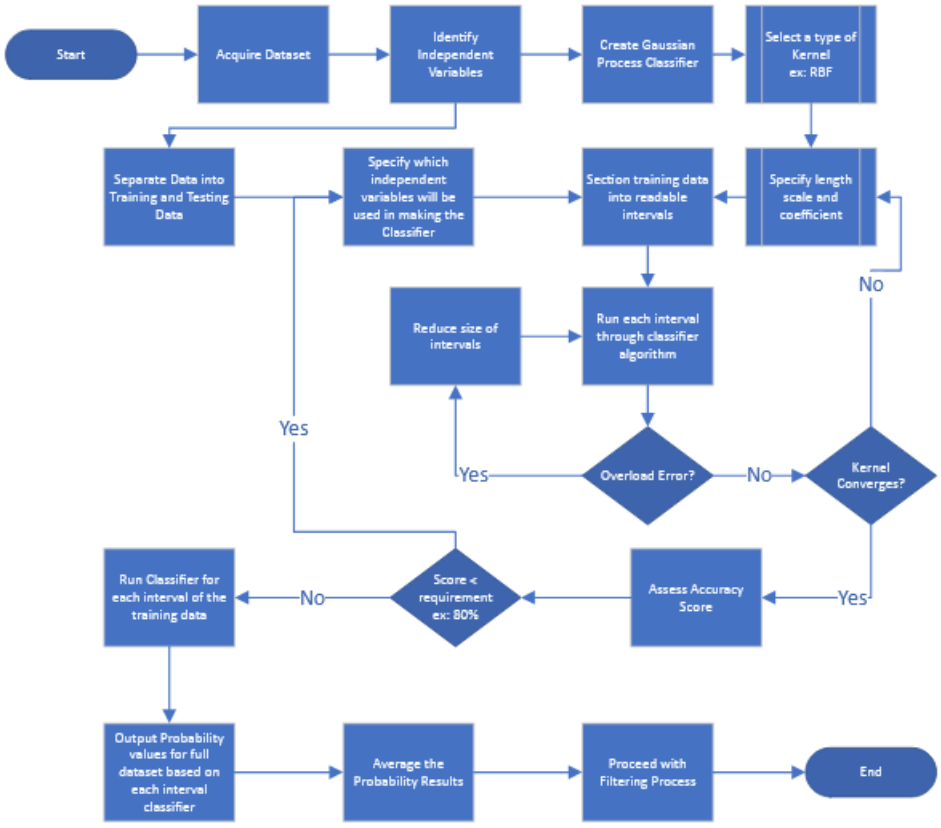

Figure 3.2: Step by step layout of the workaround created for the GPC Ensemble in order to interpret high quantities of data that the software cannot process all at once.

in this case 500 samples, and average the probability metrics outputted by each. The number of these sections is dependent on how large of a training data set the user desires; for instance, if 10,000 training data points are desired, then with an input sample size of 500 that creates 20 runs through the Gaussian process classifier. The intention behind this method is to mitigate over-fitting to a subsection of the data that has limitations on how well the overall pattern is represented.

Given the available disposable tools, this is deemed a sufficient approach and arguably a more accurate result than one that could be attempted by running substantially larger quantities of data through the classifier algorithm with the same tools. This workaround is commonly referred to in established study as an ensemble of Gaussian process classifiers, or in other words a GPC Ensemble.

One technique that was used in this study was a random sampling of 500 data points without replacement out of the total number of allotted training points that would be continuously selected at that volume until all of the training data points were now grouped in randomly selected subsets of 500 points each. The intention behind this was to encapsulate the full extend of the data without either missing any points or having points repeat in subsets too often. However, as future work it would be worthwhile to sample with replacement so as to include a form of bagging in order to cross-reference the data points within each classifier within the ensemble in order to create a sort of blending effect between them.

Additionally, the number of independent variables used when creating the classifier ensemble creates a major difference. This presents the very reason why the temperature variable was not originally incorporated into this study as was previously mentioned. Theoretically, one would imagine that the more variables used in a study, the more controlled factors identified and the more of an explicit function can be established. However, that is not found to be the case in this instance. Instead, the observation is that the accuracy score from using the classifier that incorporated temperature was significantly lower than that which only incorporated the 4 material compositions. This presents a relatively unique challenge, especially since from only implementing the classification process the user would generally not be able to understand why that is happening without

investigating the data.

Fortunately, the Sectioning and Proportioning method used in Chapter 2 does indicate a possible explanation for why that occurs. One significant observation was that when the sectioning process was done using only the 4 composition variables, there were a substantial number of intervals within which the proportional measurement was 100%. In other words, all of the data points within those intervals were deleterious due to the material compositions. That means within those regions, temperature has virtually no effect on the presence of deleterious phases. Therefore, conditionally, within those regions there would be no correlation between temperature and the probability of deleterious phases, there is no pattern there, just complete randomization. Because of that, when combining that data within the fully deleterious regions with the rest of the data, where temperature has been observed to affect the presence of deleterious phases, it becomes really difficult to model the relationship between temperature and the probability of a deleterious phase. Hence, whenever it is incorporated, the output classifier is found to be low scoring and primarily shows a random assortment of data points. This can be observed in 2-dimensional projections with respect to each material in Fig 3.3 and Fig 3.4. As can be seen, the plot without temperature incorporation has relatively good separability of the data and clear pattern, whereas with the temperature incorporation there is no clear pattern and a near completely random assortment.

While this is an observation from attempting both with and without temperature, and observing the results, for repeatability purposes in different applications it is important to establish a convention for how to select specific independent variables to use in a classification study. The best way to run this is to run through every possible number and combination of independent variables into the classifier algorithm and find which configuration produces the highest scoring classifier based on the same metric established in Equation 3.5.

3.3 Filtering

The first approach used when working with the classifier was to filter the data based on regions that have high probability of failure. The advantage to this approach is to establish a more direct specification of the variable ranges that lie within an optimal decision region rather than a more

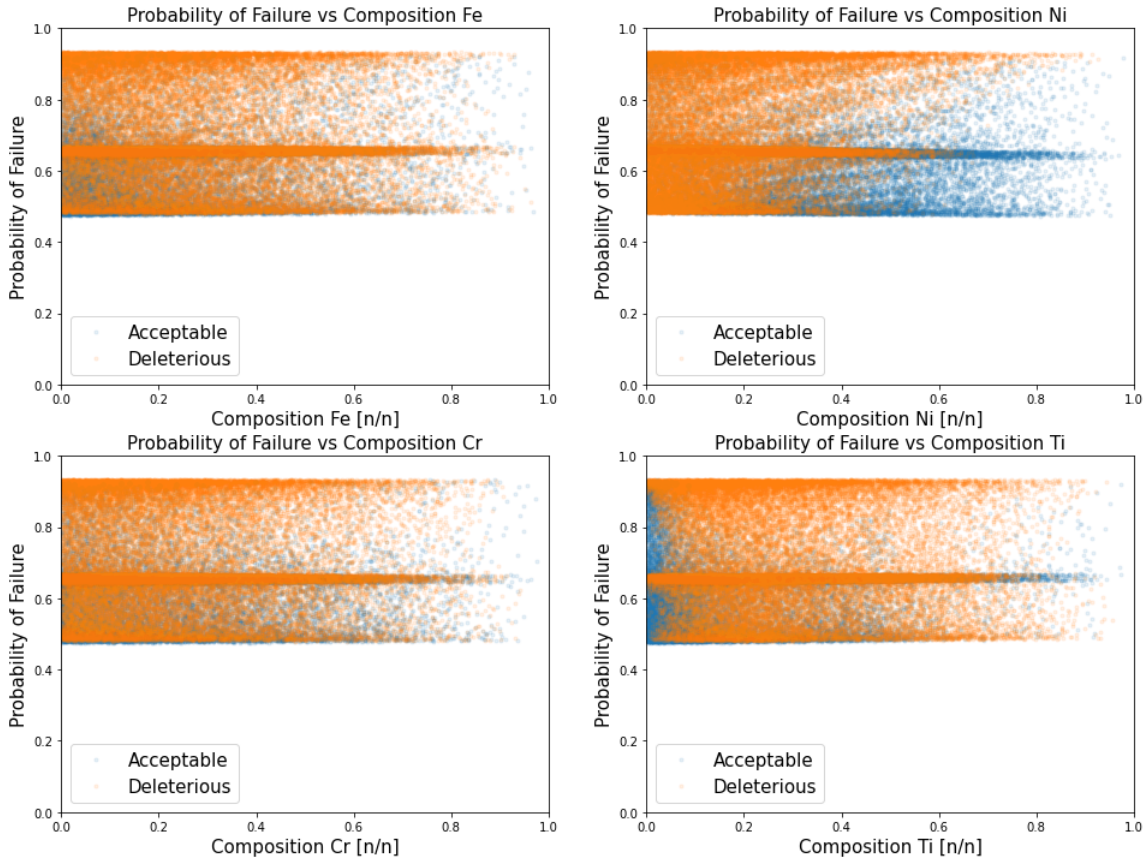

Figure 3.3: 2D Plots of the output probability of failure from the GPC ensemble with respect to each material composition along with their original labels: Blue = Acceptable, Orange = Deleterious. This Classifier ensemble was created with the 4 compositions and temperature incorporated.

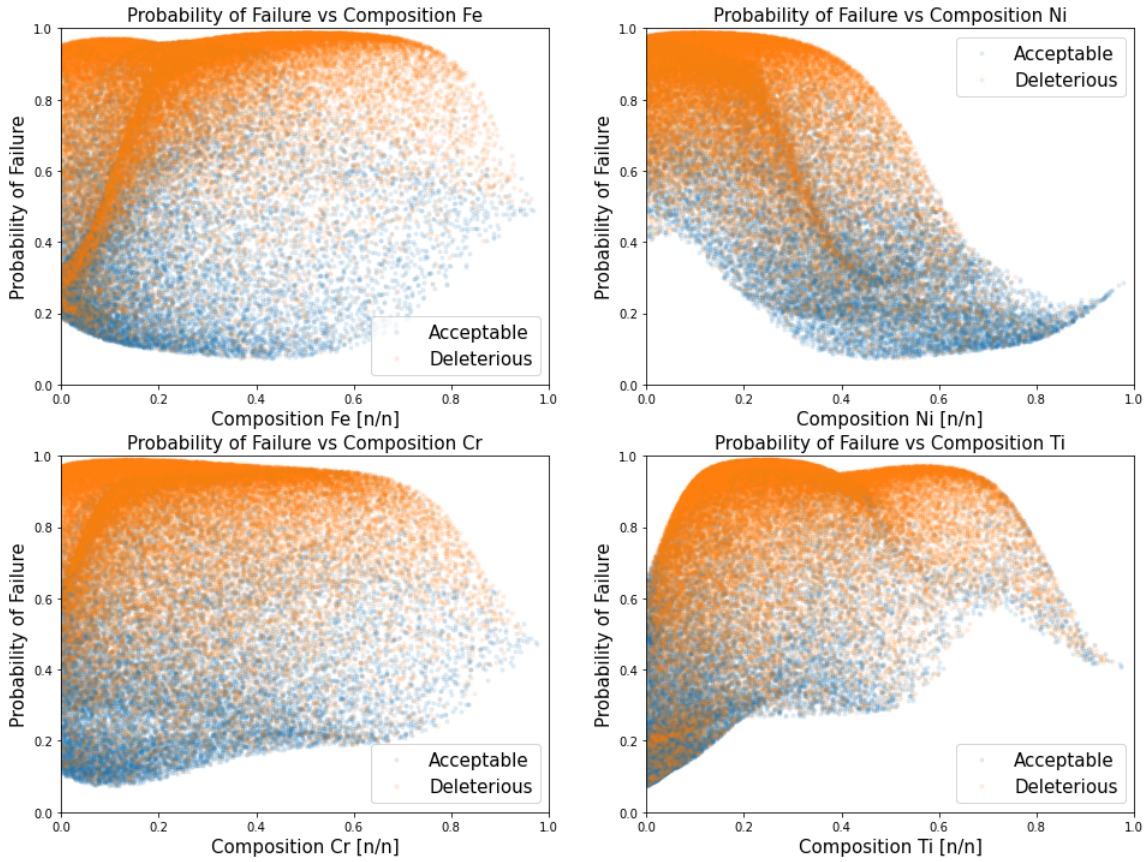

Figure 3.4: 2D Plots of the output probability of failure from the GPC ensemble with respect to each material composition along with their original labels: Blue = Acceptable, Orange = Deleterious. This Classifier ensemble was created with only the 4 compositions incorporated.

automated process produced directly by the Gaussian process classifier, which essentially means that the only presentable results exist by essentially creating a classifier ensemble in which the user would input any new testing points for which an output condition would be produced. Part of the desired deliverables of this study is to more explicitly define a range of possible values that would satisfy the intended purpose. With one or two dimensions it is relatively easy because the user could either specify a single one-dimensional variable range that works, or a combination of variable ranges that work in two dimensions. With increasing dimensionality this output becomes a lot more complicated both in terms of the ability to compute it but also in terms of the ability to interpret it.

Using a One-Dimensional approach to data filtering is arguably the easiest way to get started and understand this process, the only trade-off is that the output is more limited. By looking back at the right plot in Fig. 3.3, there is decent separability of the two conditions but for each distribution it is difficult to observe any explicit trends between the concentration of deleterious phases and each individual material composition, with the possible exception of Nickel, albeit slight. In any case, a new convention has to be applied in order to properly assess the available data. Using a similar sectioning approach as was previously presented in Chapter 2, the data in specified intervals could be averaged and projected as a more explicit trend.

3.3.1 Averaging

In order to establish a more clear trend between each independent variable and the distribution, an approached is used that is very similar to the sectioning and proportioning approach described in Chapter 2. The only difference in this approach being that instead of finding the proportion of deleterious phases within each interval, all of the probability values that exist in each interval are averaged in order to create an expectation of what the probability would be under the variable conditions within each interval[12]. The number of intervals follows a similar convention where they need to be small enough to assess the probability conditions sufficiently, but large enough such that they don't converge onto a section with too few data points. This can be observed in Fig. 3.5.

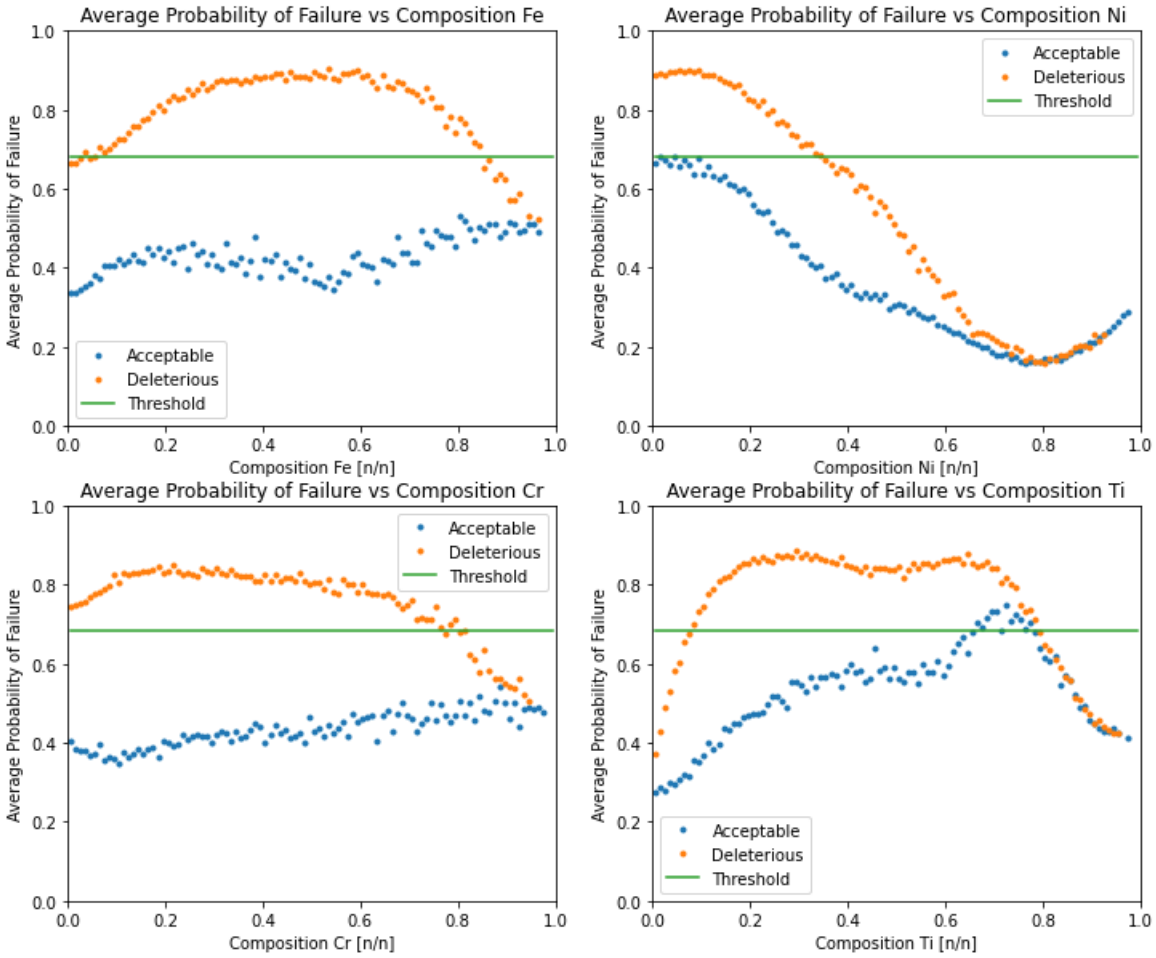

Figure 3.5: 2D Plots projecting the average probability in each interval given 100 1-dimensional interval sections with respect to each of the 4 independent composition variables. The oppositely labeled data were averaged separately to observe any differences in trends with blue = acceptable and orange = deleterious. The green horizontal line represents the total proportion of deleterious phases.

In order to see any differences present in the probability value distributions between the deleterious and non-deleterious labeled data, the full data was split into those two groups with respect to their original label with the average probability measurements being found for both cases. This was primarily done in order to see if one group's expected probability had a different relation with each composition than the other. One observation of this process is that the probability of failure for the samples that were labeled acceptable appeared to stagnate more while the probability of failure for the samples that were labeled deleterious appeared to display more of a pattern with respect to each composition. The next step was to determine the filtering criteria.

3.3.2 Criteria

The criteria for data filtering essentially represents an assessment of what factors cause deleterious phases. In a one-dimensional search, this is essentially looking for ranges of compositions that are deemed to cause deleterious phases and cut them out of the data, producing a filtered dataset that has that source of a cause of deleterious phases eliminated. There are a number of different ways that this problem could be approached. The first, and easiest, approach to use is one-dimensional filtering, where explicit ranges of each independent variable are specified in the filtered region. This process can be accomplished with either one filter region at a time or all at once. The result of one feature filter being applied with respect to composition Nickel can be displayed in Fig. 3.6.

The way these regions are specified is by condition. As can be seen in Fig. 3.5, there exists a green horizontal line that represents the total proportion of deleterious phases in the existing dataset. This is established as the filtering criteria, for which any range of averaged data with a probability expectation exceeds the total proportion of deleterious phases is filtered out, since in that case it is deemed to be a source of deleterious phases. The reason this was deemed acceptable is because the convention can be used iteratively as opposed to being user-defined each time. This filtering process can be completed using one material at a time and iteratively in succession where the total proportion is reassessed for each iteration and reestablished as the new filtering criteria. One important factor to take into consideration here is that the order in which the material

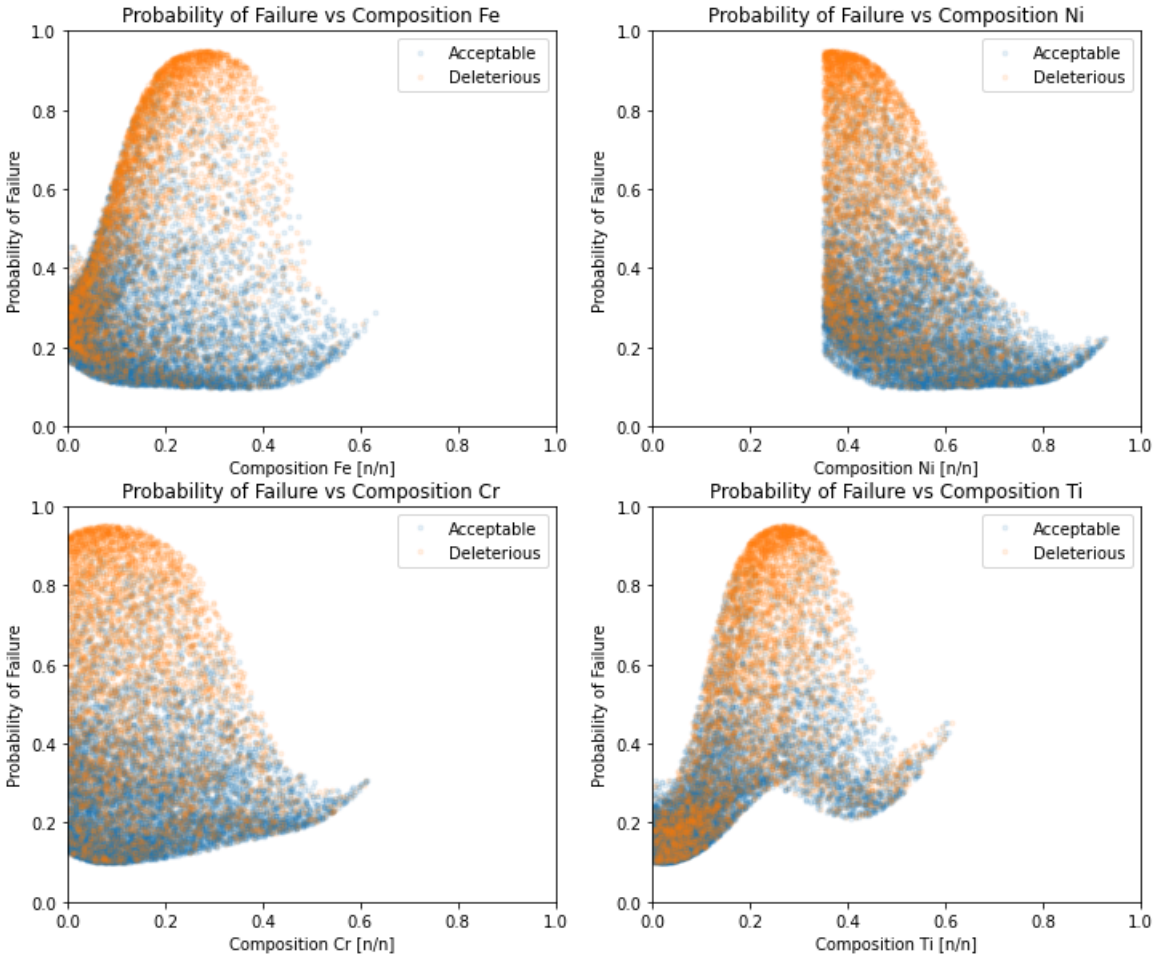

Figure 3.6: 2D Plots projecting the filtered data points with respect to Nickel based on the previously established criteria from Fig. 3.5 probability in each interval. The data labels are consistent with blue = acceptable and orange = deleterious.

composition is filtered makes a difference, because as sources are filtered with respect to one composition, the results that are projected with respect to the other compositions will change, particularly the overall probability of failure that exists in each composition.

Therefore, when the averaged probabilities per interval are computed for a second time, they will differ from that of the first because a section of the points have been filtered out, which changes the projections across each of the other independent variables. Therefore, because of this, the order in which compositions are filtered makes a difference. The only way to know which will produce an optimal result, which includes a region with a balance between lowest proportion of deleterious phases in the filtered data and highest number of points within the filtered data, is to run through every possible order of composition filtering and compare the results. The challenge with that approach is that running through each possible order takes a substantially long time to compute and therefore is more impractical in comparison to alternatives. Alternatively, all 4 features can be filtered at once to arrive at an outcome that is perhaps not the optimal result but can be achieved the quickest.

3.3.3 Higher Dimensions

While running this method as a one-dimensional practice is the easiest way to understand how the filtering concept works, it is definitely far from the best process to use because it is very limited in its ability to determine conditional probability estimations. While the averaged probability value per interval works sufficiently well at estimating how the probability of a deleterious phase being present will change with respect to each independent variable, it still maintains high variance which means that samples do exist within those filtered regions that are still usable under other conditions. However, those conditions cannot be found using a one-dimensional filtering approach. Therefore, the intention is to expand the filtering process into multiple dimensions, where this time instead of being one-dimensional sectioning, averaging, and filtering it instead becomes a grid search for spaces that contain an average probability that exceeds the total proportion of deleterious phases.

3.4 Decision Region

Switching objectives, while the intention of including the filtering process, in theory, is practical from the perspective of user interpretation of being able to assess conditions of a decision boundary directly, the result of an optimal decision boundary can be assessed much more quickly and easily just through applying the classifier ensemble directly and filtering the data based on the probability outputs from that, where any point that is classified as having greater than a 50% chance of a deleterious phase is filtered out. This automatically interprets conditional decision regions based on one another, which eliminates the problems with the filtering procedure of both neglecting those conditional regions and taking a high computation time. However, the decision region, while optimal, still contains a level of uncertainty that would be insufficient for a practical application. This goes back to the original classifier setup where temperature was not originally incorporated because of how it resulted in a poorly fit classifier.

In order to further converge onto an optimal decision region with minimal uncertainty, there are two approaches that can be used. The first involves setting the filtering criteria from the probability outputs with respect to a lower threshold. However, a more practical approach, which incorporates the previously uninvestigated variable of temperature, should be incorporated into the next steps.

3.5 Temperature Dependency

The way the temperature measurement factors into the output of this study is not consistent and therefore difficult to project adequately. As was shown from the sectioning and proportioning approach in Chapter 2, there are a substantial number of regions bases solely on the 4 material compositions within which the probability of a deleterious phase being present is estimated to be 100%, which means that regardless of what value temperature is within those regions, the phase will always be outputted as a deleterious phase. Hence, temperature has no effect on the output within those regions, which means that any trend between those two variables in those regions would be completely random. However, as explained in Computational Design of Compositionally Graded Alloys for Property Monotonicity[5], thermal expansion is one of the factors in play that

Proportion of Deleterious Phases Per Interval

Temperature	Interval Range: 0.2 < Fe < 0.4, 0.6 < Ni < 0.8, 0 < Cr < 0.2, 0 < Ti < 0.2	Interval Range: 0.2 < Fe < 0.4, 0.2 < Ni < 0.4, 0 < Cr < 0.2, 0 < Ti < 0.2	Interval Range: 0.6 < Fe < 0.8, 0 < Ni < 0.2, 0 < Cr < 0.2, 0.2 < Ti < 0.4
300K < T < 450K	0.17948718	0.36585366	1.0
450K < T < 600K	0.28205128	0.57894737	1.0
600K < T < 750K	0.63043478	0.60465116	1.0
750K < T < 900K	0.65853659	0.91891892	1.0
900K < T < 1050K	0.71794872	0.92857143	1.0
1050K < T < 1200K	0.86486486	0.87096774	1.0
1200K < T < 1350K	0.69047619	0.86666667	1.0
1350K < T < 1500K	0.86956522	0.98	1.0
1500K < T < 1650K	1.0	1.0	1.0
1650K < T < 1800K	1.0	1.0	1.0

Table 3.1: Table measurements of proportions of deleterious phases per interval with respect to 3 specified composition ranges and temperature ranges.

will influence the properties and phases present, meaning that there has to be some cause between increased temperature and undesirable properties.

Therefore, through the use of the sectioning and proportioning approach from chapter 2 where ranges of temperature values are established in intervals and projected with respect to each previously established interval based on the 4 compositions, the observation displays a decrease in the overall proportion of deleterious phases within smaller temperatures of regions within which the total proportion of deleterious phases are less than 100%. An example of this involving 3 specific intervals across multiple temperature ranges can be observed in Table 3.1 This indicates a conditional secondary effect that temperature has on the output, where it only changes the result under certain conditions. Therefore, this has to be incorporated.

When implementing the approach within the context of the GPC ensemble method, there are two main approaches used. The first involves a one size fits all hard cutoff temperature which is optimally spaced such that the false positive and false negative errors are minimized using Bayes Theory[3]. The second involves implementing a second GPC ensemble to be implemented after the first which incorporates temperature.

3.5.1 Hard Cut-off

The hard cutoff approach is relatively easy to implement given that, at its core, the procedure can be as simple as the user specifying a maximum allowable temperature value and filtering the rest of the data accordingly, such that all data points with a temperature value above that maximum threshold are filtered out. Optimizing this to minimize error from both sides, however, can prove to be a challenge. Therefore, it is worthwhile to incorporate a form of Bayes theory in this case. Recall from Equation 3.1 the approach of multiplying the prior probability with a likelihood function. In this case, both of those quantities have to be identified with respect to the data.

Starting with the prior probability, this can be assessed using the same proportioning approached used previously and applied to the two specific regions that exist both above and below the temperature threshold. In other words, it essentially presents a conditional probability that states "The probability of a deleterious phase existing is P(E), given that the condition of the points within that subset all meet the criteria of being below the specified temperature threshold". The likelihood function, which is applied to the prior probability to create a posterior probability estimate, is represented by the proportion of samples that fit on either side of the specified cutoff temperature. This value will scale with respect to the value set at the cutoff temperature.

In order to find the optimal value, a sufficient approach involves doing an iterative search to minimize the expected error. This approach is described in a step by step format in Fig. 3.7. To visualize how the error is minimized more directly, a plot representing the posterior probabilities and the area under the curve shown as the error can be seen in Fig. 3.8

3.5.2 Second Classifier

This process is very straightforward, it essentially is a repetition of the original classification process with the incorporation of temperature into the classifier algorithm. This second classifier ensemble is applied to the optimal decision region from the original classifier ensemble. This time it includes all four material compositions and the corresponding temperature, as opposed to originally where it just consisted of the four material compositions.

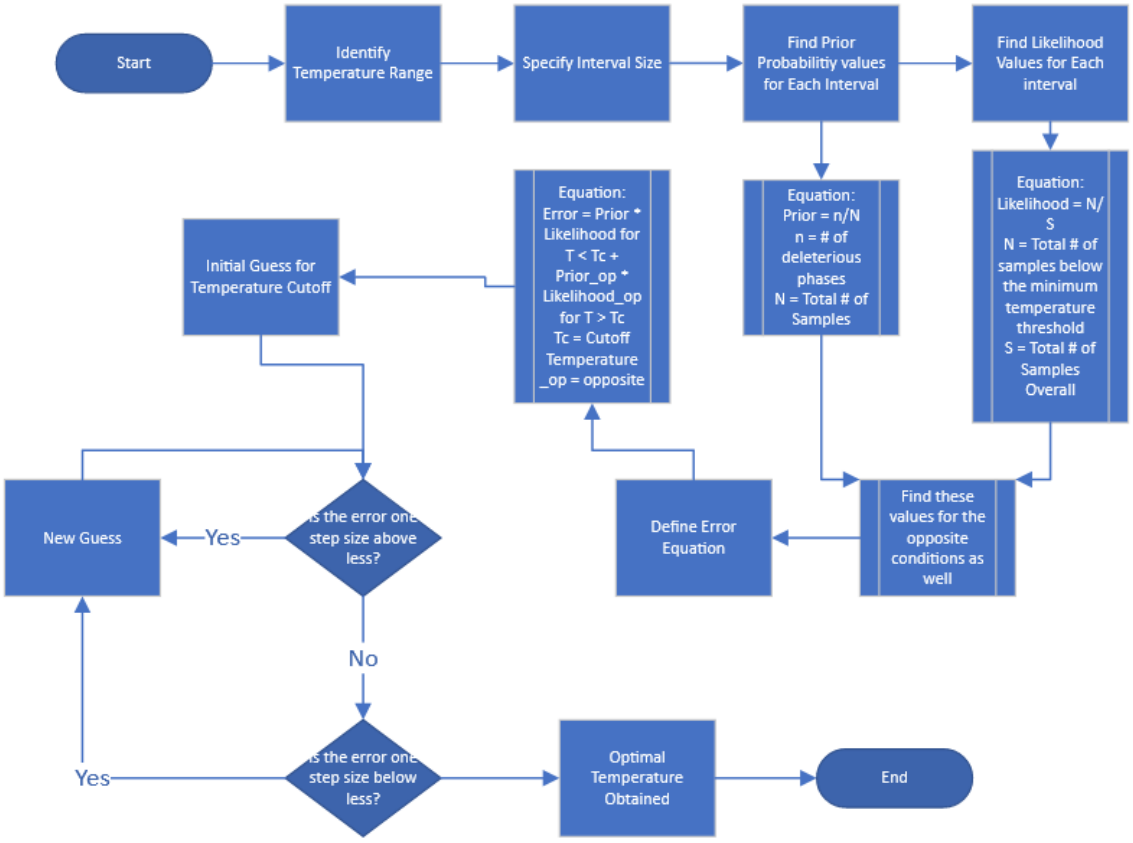

Figure 3.7: Step by Step Process displaying how to find the optimal cut-off temperature using Bayes Theory.

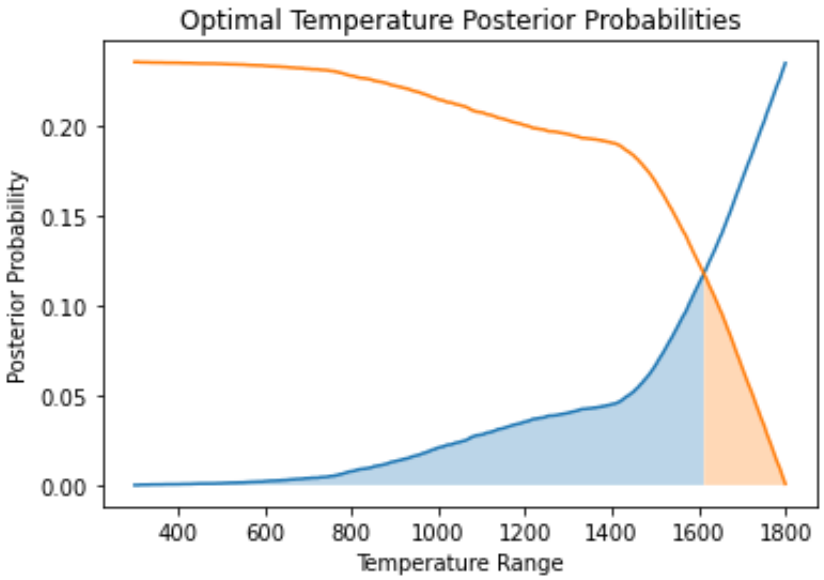

Figure 3.8: Optimal Error found by iteratively searching for a temperature value for which the false positive and false negative values would be minimized.

The reason why it is able to account for the temperature in this instance when it wasn't before can be theorized as having to do with the observation found in Chapter 2 with the sectioning and proportioning method where for composition ranges that contained 100% deleterious phases, the temperature would have had no effect on the output. However, since with the original GPC ensemble, that combination of material compositions would have been filtered out, which could explain why temperature is successfully integrated into the second round GPC ensemble.

This is able to converge onto an optimal decision region that produces less of a false positive error than the hard cut-off temperature method because it incorporates conditional cases where the max allowable temperature would change with respect to each material composition. However, while it does produce a region with the most minimal uncertainty attainable from this study, it still has enough uncertainty particularly around the decision region such that there still exists some probability of failure. To lower it further, more methods will have to be employed, or the probability metric outputted by the classifier will need to be filtered with respect to a lower allowable threshold.

4. CONCLUSIONS AND FUTURE WORK

4.1 Results

In conclusion, both of the methods under analysis in this study have their advantages and disadvantages and, as future work, could be more integrated with one another. The most significant benefit to the sectioning and proportioning method involved a more direct investigation and measurement of exactly how each variable affects the output, which in this case is how likely a deleterious phase is to exist under certain conditions. Whereas applying a classifier directly is able to more thoroughly and with greater certainty predict a probability estimation, it does not take those aforementioned factors into as much consideration and it is also difficult to infer any extent of uncertainty in the predictions.

When comparing the results of the two classifiers, surprisingly the results of the approach that involves a Gaussian process regression of the proportional measurements of the sectioned data is closer than anticipated to the estimate found via direct Gaussian process classification. With further development of this approach and testing in various applications, a convention with the sectioning and proportioning approach could be found to have substantial validity.

4.1.1 Sectioning and Proportioning

The most substantial takeaway from this study, which could not be inferred from the classifier approach, was the ability to identify the exact scope and thoroughness of the existing data as well as identify the way that data is distributed throughout the sections. This is a factor that is very difficult and often near impossible to identify through just inputting a subsection defined as training data into a Gaussian process classifier. It is also very easy to apply, it is a relatively simple approach of just taking a fraction of deleterious phases over total number of phases and applying that fraction to intervals of the data defined by variable ranges.

The process of defining interval size and reliability is very useful for identifying how much information is present at any given location of the data. The overall distribution of data within

the sample space is something that can often be overlooked especially with high quantities of samples like with this dataset containing 50,000, that said it is able to identify areas of limited knowledge or understanding, especially when there is already implied uncertainty to begin with. In this dataset it can be seen through this kind of search that the data distribution is most certainly not uniform throughout the entire sample space. There are numerous regions that are found to have a greater sample point density than others. This, in essence, corresponds to the amount of confidence that can be had in the predictions. After all, a generally acceptable convention is that the more information present in any given case, the more confident are any conclusions that can be drawn from it.

The implementation of a Gaussian process regressor to generate a probability metric that exists as more than a step function and also incorporates uncertainty in the predictions does still have a ways to go in order to prove to be a viable approach, which is a strong argument in favor of using the classifier approach in this instance. One of the most significant factors that is relevant in this instance is error estimation, which can make a significant difference in the generation of the surrogate model, and without a ground truth to compare the results to it becomes difficult to assess in terms of its validity, especially when projected across multiple dimensions.

However, one of the weaknesses of the Gaussian process regressor, which primarily has to do with the error estimation, is that based on the data distribution it cannot always sufficiently predict a surrogate model that properly encapsulates the scope of the data as seen from the proportional measurements. If there exists a higher point density around a particular output and the error estimation is loose, then the surrogate model is more likely to interpret the data points with less density as erroneous or noise, and the result will be substantially offset from them. The way to accommodate for this is to have a lower error estimation, however that requires the assumption that the data at that point is more certain without any way of verifying. An example of this occurrence is shown in Fig. 4.1 and Fig. 4.2, where two surrogate models are generated with the same parameters and based on the same type of data with the exception of one has more samples than the other.

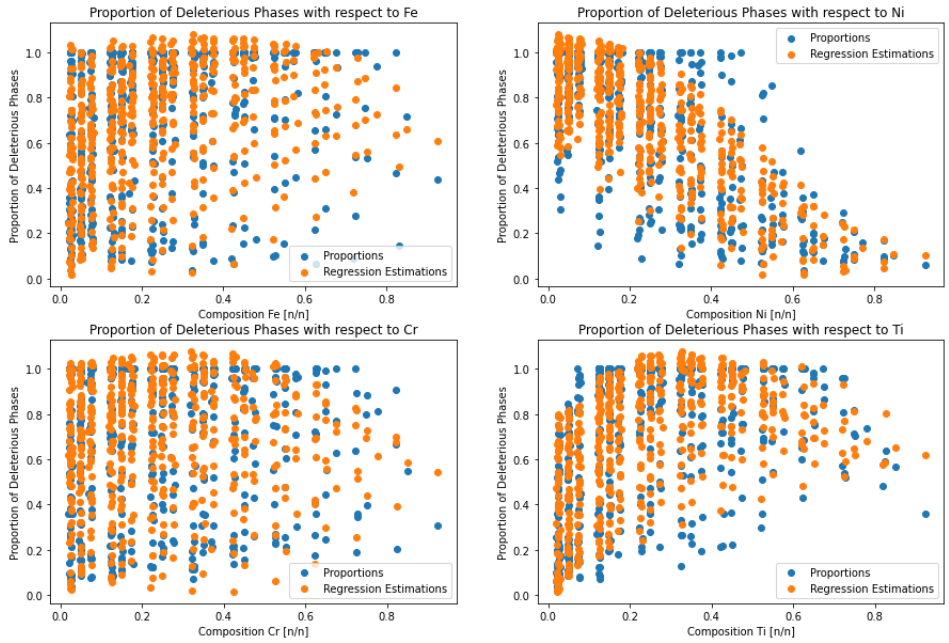

Figure 4.1: Proportional Measurements per interval compared alongside their corresponding regression estimations. Blue = Proportional measurement, Orange = Regression Output.

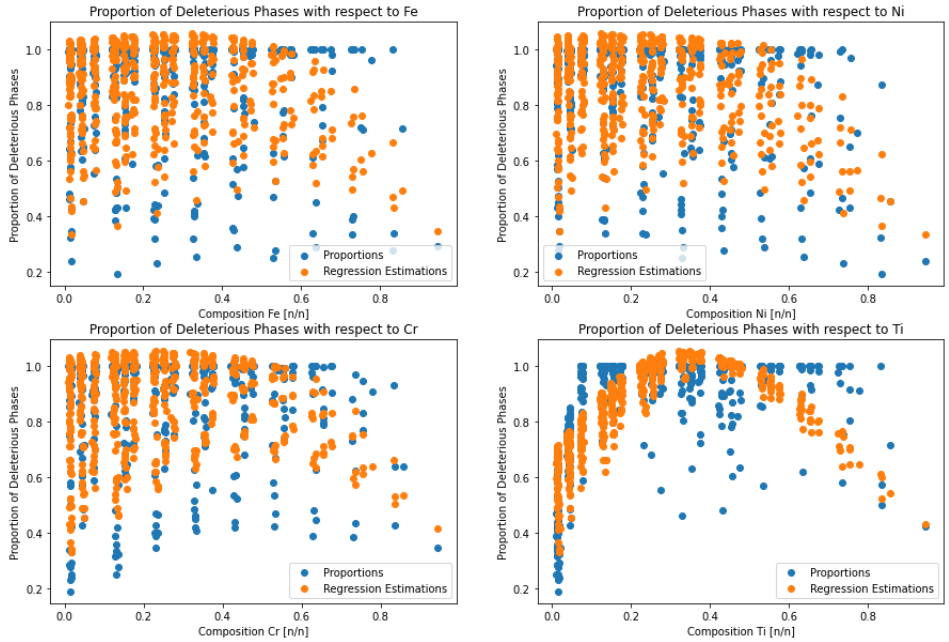

Figure 4.2: Proportional Measurements per interval compared alongside their corresponding regression estimations. These plots contain a greater quantity of sample points that are more heavily concentrated in certain areas of the sample space than the ones in Fig. 4.1. Blue = Proportional measurement, Orange = Regression Output.

4.1.2 GPC Filtering

The disadvantages of this section significantly outweigh the benefits mainly because of the substantially long computation time it takes to run thoroughly in comparison to the validity of the results. The one significant benefit that comes from using this approach is providing a more direct user-interpretable version of the results as opposed to setting it as an automated process for which assessments over possible material and temperature combinations of samples would need to be run through the classifier and assessed directly by the machine to determine its viability. However, the disadvantage here is that when many conditional cases are assessed such that the identified viable design space is maximized, so many dimensions are used and so many conditional variable ranges are present such that it makes more sense to automate it because the user-interpretability decreases substantially with more thoroughly established decision regions dependent on multiple variables.

This was the first approach used in this study, hence why it is more imperfect than newer versions. That said, the most valuable lesson that came out of this study was the idea of taking intervals and finding probability averages within to establish an expression for what the expected probability of a deleterious phase is under certain conditions. This idea is what led to the sectioning and proportioning approach as a more direct method without the prior incorporation of a classifier algorithm and the research, experimentation that has involved that. This method is what inadvertently set the stage for an approach that, with further study and experimentation and perhaps expansion into other applications, could have the potential to be an effective approach at solving these kinds of problems. The iterative filtering approach based on a Gaussian process classifier output, however, appears to be at a dead end in terms of its feasibility. The concepts of data filtering and sectioning, however, merit further incorporation and investigation.

4.1.3 Temperature Dependent GPC

This approach is, at this stage of the study, the most effective and straightforward approach when it comes to arriving at a viable solution. The main observation from this stage, however, is that there is enough uncertainty around the decision boundary such that it becomes difficult to

Classifier Results

	Original	Post Classifier	Post Temp Classifier	Post Hard-Cutoff Temp
Proportion of deleterious phases	68.3%	23.6%	8.8%	13.5%
Number of Data Samples	50,000	14,370	11,600	12,573

Table 4.1: The results of each of the classifier ensembles applied in terms of the proportion of deleterious phases in the existing data as well as the number of samples present in each version of the data.

infer a large decision region with minimal uncertainty based solely on the classifier itself. Because of this, the optimal decision region still contains a substantial amount of false positive error that, in an ideal circumstance, would be reduced further. The original data metrics along with the corresponding probability estimations are shown in Table 4.1.

In order to further reduce this false positive error, the filtering criteria based on the probability of failure has to be reduced to gradually lower amounts. The same can be done for the original classifier as well. The trade off with that is that the usable feature space is further reduced. However, for the purpose of reducing the probability of failure to below a desirable threshold such as 1%, that would be the most viable approach given this method in its current state.

4.2 Comparison of Results

By using the Gaussian process regressor established in the sectioning and proportioning approach and applying it to the full dataset, a projection can be obtained overall that appears to behave very similarly to the projection resultant from the Gaussian process classifier ensemble approach. While the results of the Gaussian process classifier ensemble method can still be considered the superior approach in terms of its ability to converge onto an optimal decision region, the regression of proportional measurements was seen to be an effective approach when it comes to finding a decision region and quantifying the probability. The comparison of the results of these two approaches in terms of both the number of samples and the proportion of deleterious phases after each stage are shown in Table 4.1 and Table 4.2. A more direct projection of the regression

Probability Regression Results

	Original	Post Probability Regression	Post Temp Probability Regression	Post Hard-Cutoff Temp
Proportion of deleterious phases	68.3%	23.5%	23.5% (no change)	13.4%
Number of Data Samples	50,000	12,767	12,767 (no change)	11,189

Table 4.2: The results of each of the the regression outputs of the sectioned proportional measurements applied in terms of the proportion of deleterious phases in the existing data as well as the number of samples present in each version of the data.

results can be seen in Fig. 4.3.

With some further development and implementation into other applications, this approach could be established as a usable convention moving forward. One of the key advantages that are present in this approach is the ability to investigate the scope of the data directly. The disadvantage, however, comes in the form of the fact that various conventions of this approach including but not limited to the size and number of variables used in each section of the data within which proportional measurements are taken is very problem specific and as of right now the established method of finding those parameters is through trial and error.

4.3 Next Steps

The next steps for this study primarily involves further development of the gaussian process regressor approach that incorporates the proportion measurements from the sectioning and proportioning method, particularly with respect to the error estimator. The main reason why sectioning and proportioning is deemed useful from this attempt is because the direct investigation and measurements of proportions that can be obtained from it do present useful information about the data incorporated into this study that is not obtained solely through the use of using the data to train the classifier. It would be useful to further investigate the process of regressing the proportion results into a surrogate model to represent the probability metric for which the intent of this study is to quantify.

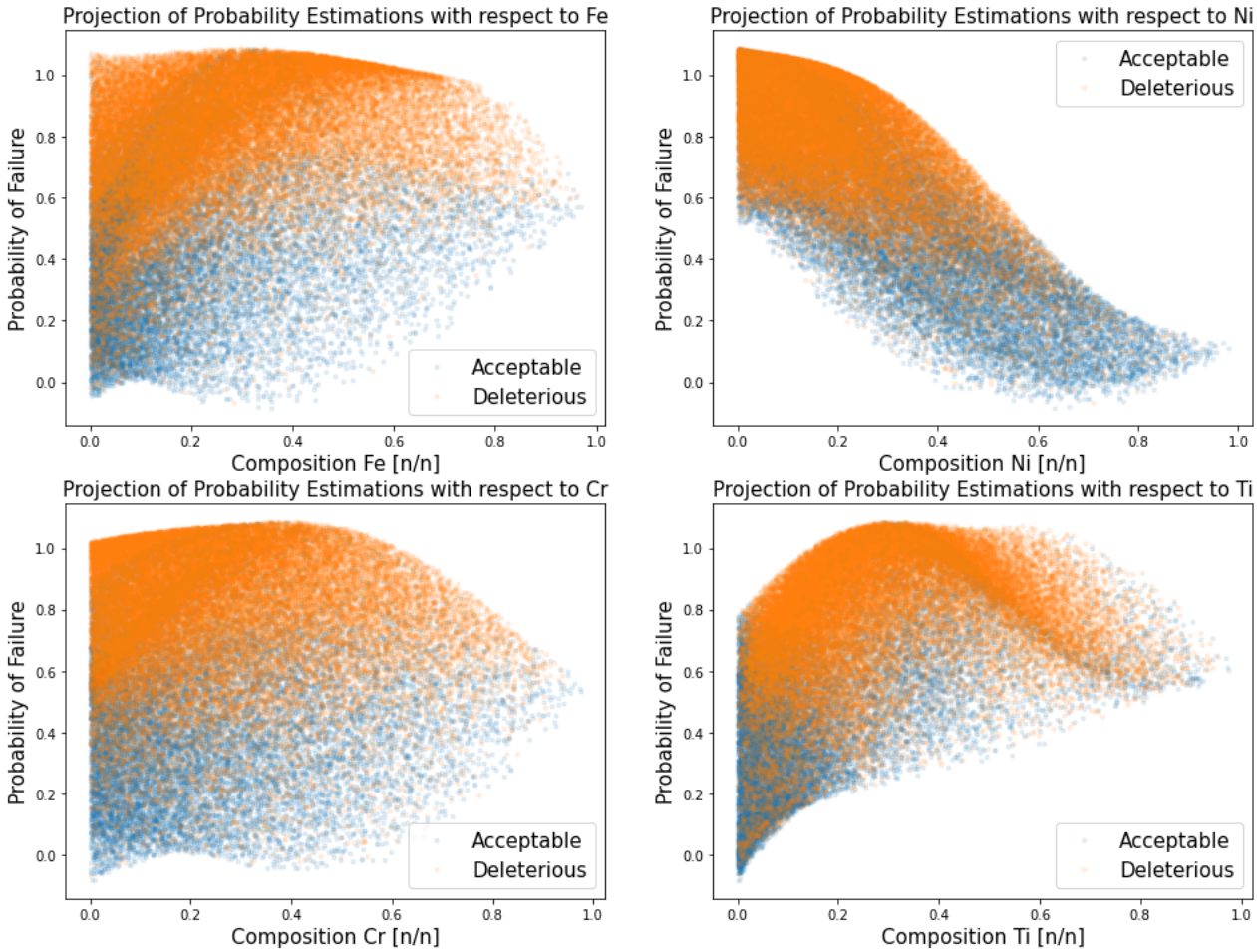

Figure 4.3: 2D Plots of the output probability of failure from the Gaussian Process Regression Output of the Sectioned Proportional Measurements with respect to composition Iron along with their original labels: Blue = Acceptable, Orange = Deleterious. Left Plot: Classifier with 4 compositions and temperature incorporated; Right Plot: Classifier with only the 4 compositions incorporated.

The hope with this is that if a general convention can be established that works not just for this application but has potential for multiple different applications then that presents a viable tool to be used for future studies not just as an alternate means of uncertainty quantification in classification problems but also optimizing to ensure that a certain condition is met. However, there still exists a ways to go to arrive at that point. The most important factor to establish is error estimation, which is still ambiguously defined and based on assumption at this point.

Additionally, as far as classification is concerned, a more thorough convention has to be established and trials using different kinds of classifiers such as Support Vector Machine(SVM) and Known Neural Networks(KNN) are worth implementing in order to compare the findings from them to the findings of the Gaussian Process Classifier and see which could generate.

As for the Gaussian Process, one other factor that wasn't heavily investigated in this particular study but would be a worthwhile secondary area of investigation would be to modify the kernel function that is used. The function that was used in this approach was a radial basis function. However, that is certainly not the only function that can be used in this case. It would be worth attempting other forms of the kernel function to observe any changes in the execution of this process as a secondary future study as well.

One of the rationales behind the methods introduced in this study is to investigate and showcase ways to interpret mass amounts of data with more commonly available tools and software. For a more in depth and thorough investigation more advanced computational software with higher processing capacity could be useful in order to not run into the problem of too much data overloading the module. However, that also depends on how exact the investigation is intended to go as well, and if there is a way to make easier to use and more readily available tools work for the intention of what is sought after to accomplish, then that may still suffice in a number of applications.

Lastly, in this instance with temperature incorporated as a secondary conditional variable, it is worthwhile to further investigate ways to incorporate this more directly. With the classifier approach a secondary step had to be taken in order to find another more optimal decision region. The disadvantage to this is that since it is already active on the original filtered region, there were many

sections filtered out which could have been found to be potentially been viable parts of the optimal decision region if temperature was incorporated more directly. This presents an additional rationale for regressing the sectioning and proportioning output, since that is based on direct measurements, in theory the output would be able to accomodate that in a way that the classifier ensemble wasn't able to initially.

REFERENCES

[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Adaptive computation and machine learning, Cambridge, Mass: MIT Press, 2006. OCLC: ocm61285753.

[2] S. A. Sudip Dey, Tanmoy Mukhopadhyay, *Uncertainty Quantification in Laminated Composites*. Boca Raton, FL: CRC Press, 2019.

[3] E. Brochu, V. M. Cora, and N. de Freitas, "A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," *ArXiv*, vol. abs/1012.2599, 2010.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer.

[5] T. Kirk, E. Galvan, R. Malak, and R. Arroyave, "Computational Design of Gradient Paths in Additively Manufactured Functionally Graded Materials," *Journal of Mechanical Design*, vol. 140, Sept. 2018.

[6] H. Mao, H. L. Chen, and Q. Chen, "Tchea1: A thermodynamic database not limited for high entropy alloys," 2017. J. Phase Equilib. Diffus.

[7] K. R. Hoffer J.G., Geiger B.C., "Gaussian process surrogates for modeling uncertainties in a use case of forging superalloys," *applied sciences*, vol. 12, 2022.

[8] C. J. H. W. C. D. Z. M. H. Q. C. Z. B. F., "Gaussian process kernel transfer enabled method for electric machines intelligent faults detection with limited samples," *IEEE*, 2021.

[9] W. W. Jones D.R., Schonlau M., "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, vol. 13, p. 455–492, 1998.

[10] E. Fink, A. Sarin, and J. G. Carbonell, "Analysis of uncertain data: Smoothing of histograms,"

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher,

M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[12] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. Hoboken, NJ: Wiley, 7th ed., 2018.

APPENDIX A

SECTIONING AND PROPORTIONING APPENDIX

The full outputs from the Gaussian Process Regression across the full sample space are displayed in the following figures. Each output is projected over each material composition in a 1-dimensional projection.

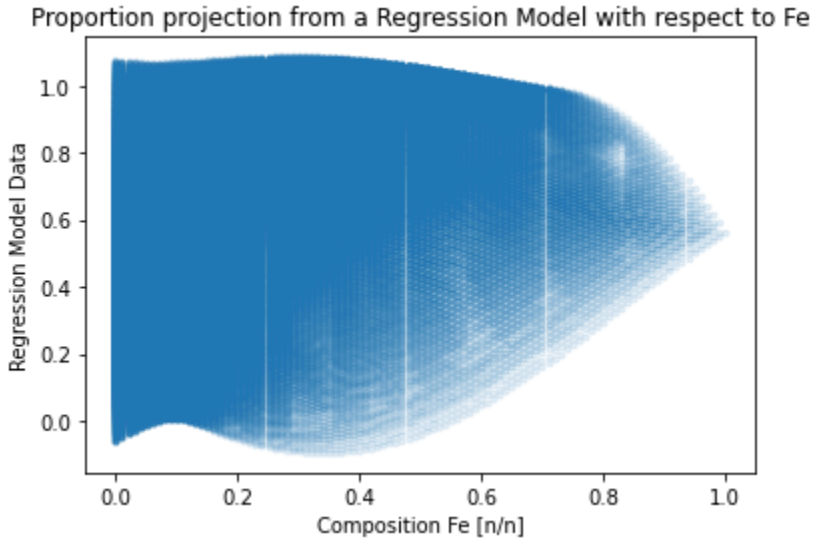

Figure A.1: Probability Estimations based on the Gaussian Process Regression over the entire sample space projected over Fe.

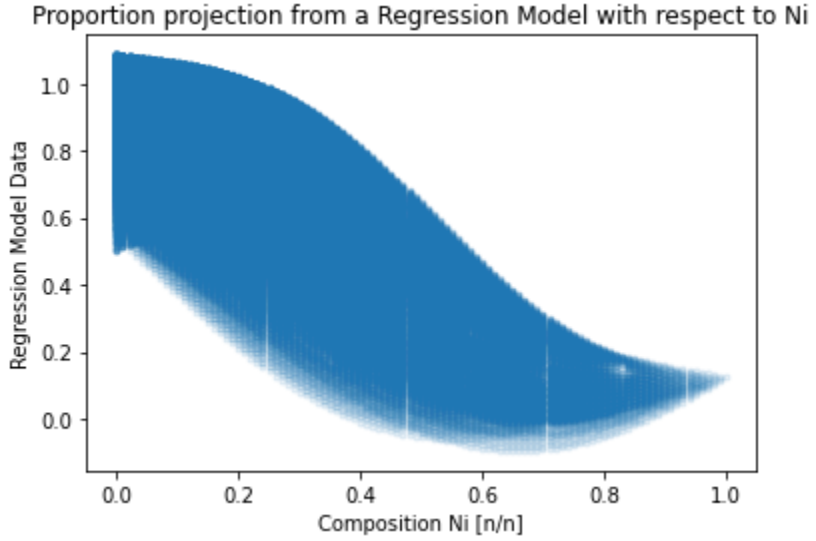

Figure A.2: Probability Estimations based on the Gaussian Process Regression over the entire sample space projected over Ni.

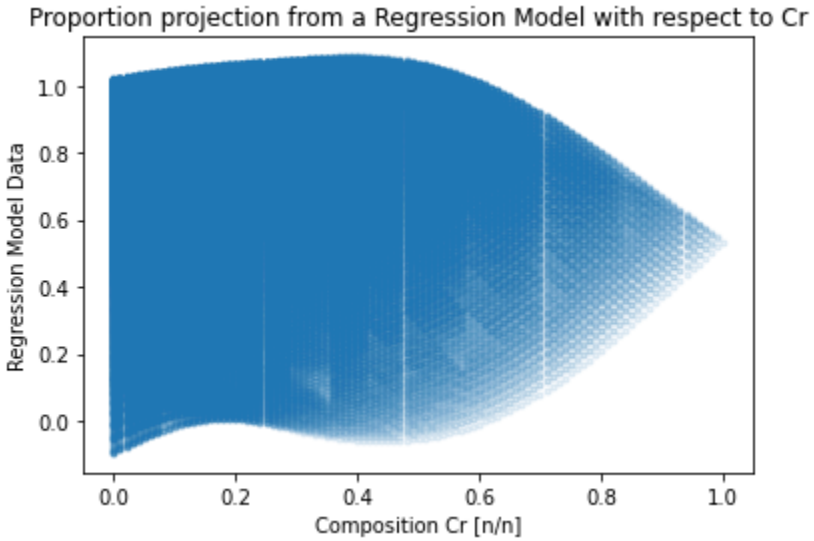

Figure A.3: Probability Estimations based on the Gaussian Process Regression over the entire sample space projected over Cr.

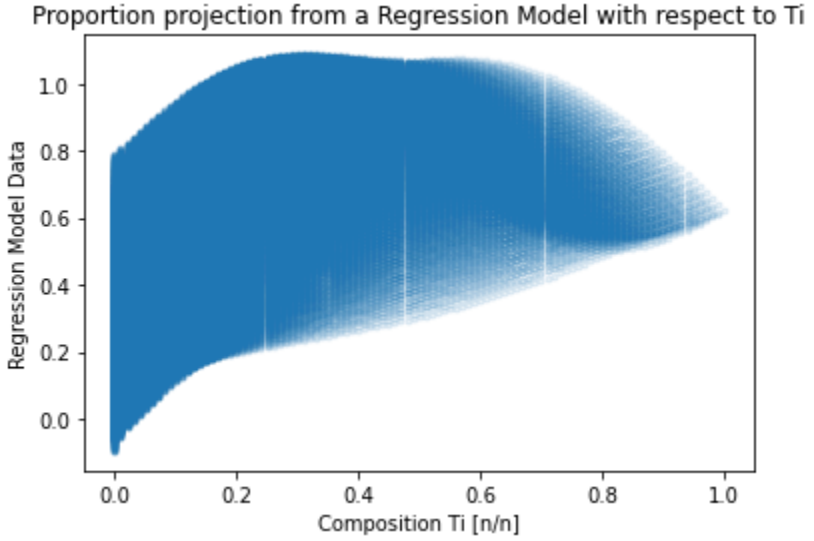

Figure A.4: Probability Estimations based on the Gaussian Process Regression over the entire sample space projected over Ti.

APPENDIX B

CLASSIFICATION APPENDIX

B.1 Pre and Post Classifier Filtering

The full outputs from the Gaussian Process Classifiers across the full sample space are displayed in Fig. B.1. Each output is projected over each material composition in a 1-dimensional projection.

The projected results of the optimal decision region found by the original classifier are displayed in Fig. B.2

B.2 Post Temperature Incorporation

The projected results of the optimal decision region found after the hard cut-off temperature threshold is applied to the filtered data are displayed in Fig. B.3

The projected results of the optimal decision region found after the second classifier with temperature incorporated is applied to the filtered data are displayed in Fig. B.4

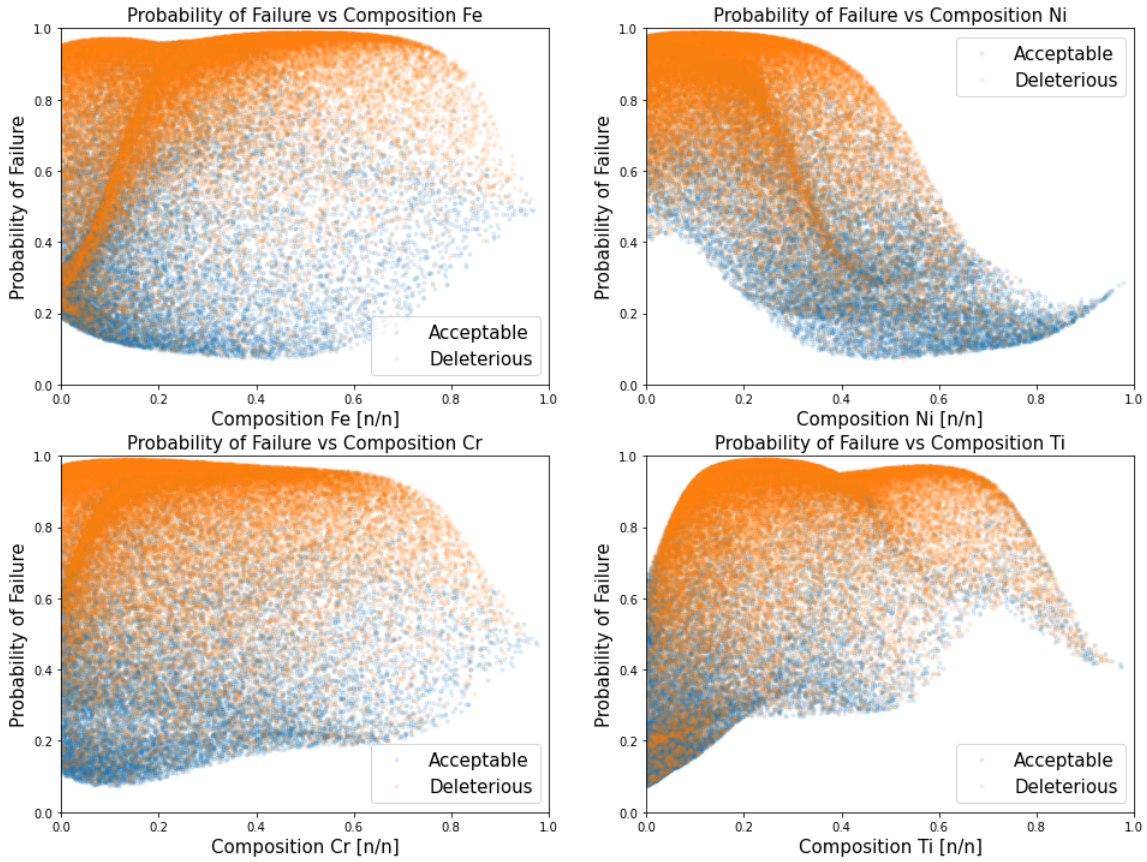

Figure B.1: Probability Estimations based on the Gaussian Process Classifier over the entire sample space projected over each material composition. Blue = Acceptable Phase, Orange = Deleterious Phase.

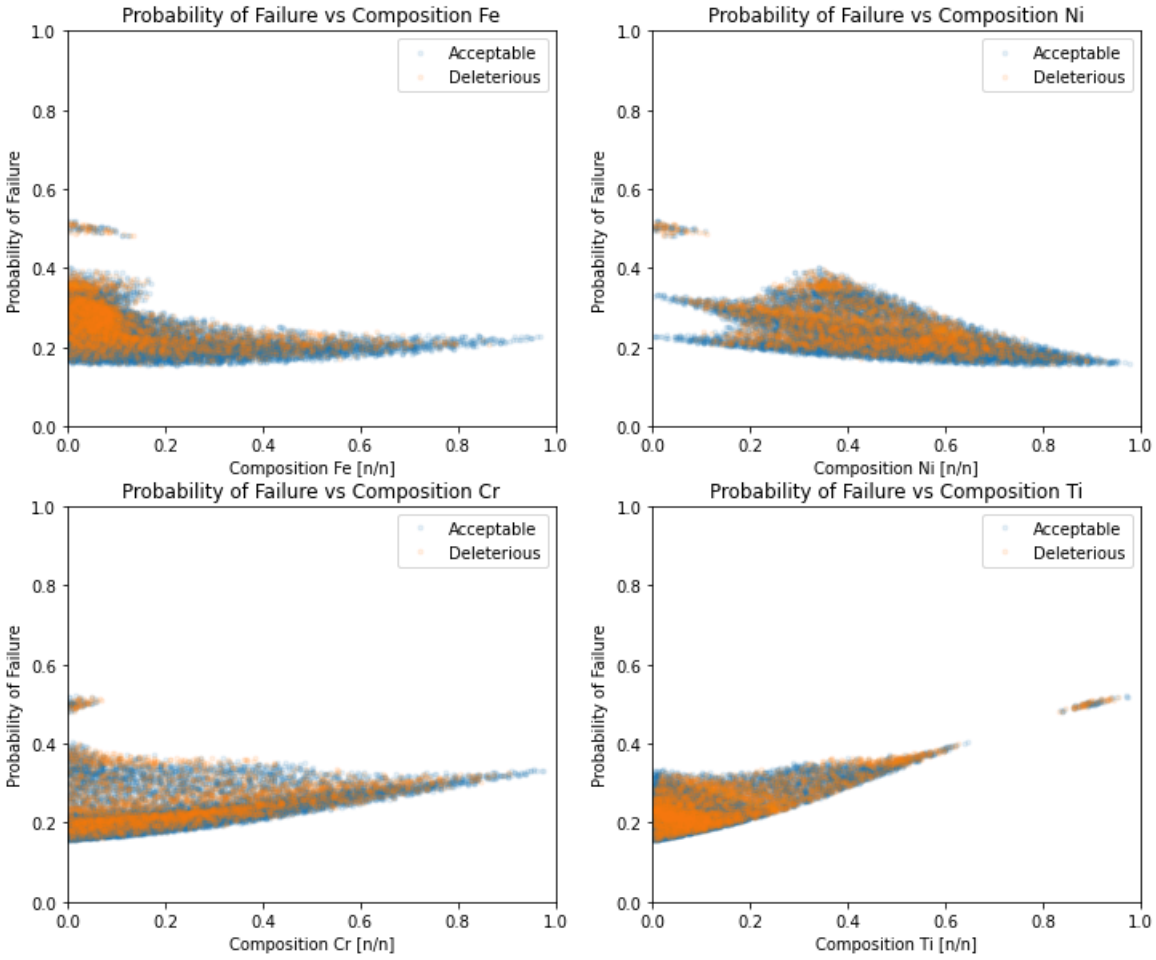

Figure B.2: Filtered Probability Estimations based on the Gaussian Process Classifier over the entire sample space projected over each material composition. Blue = Acceptable Phase, Orange = Deleterious Phase.

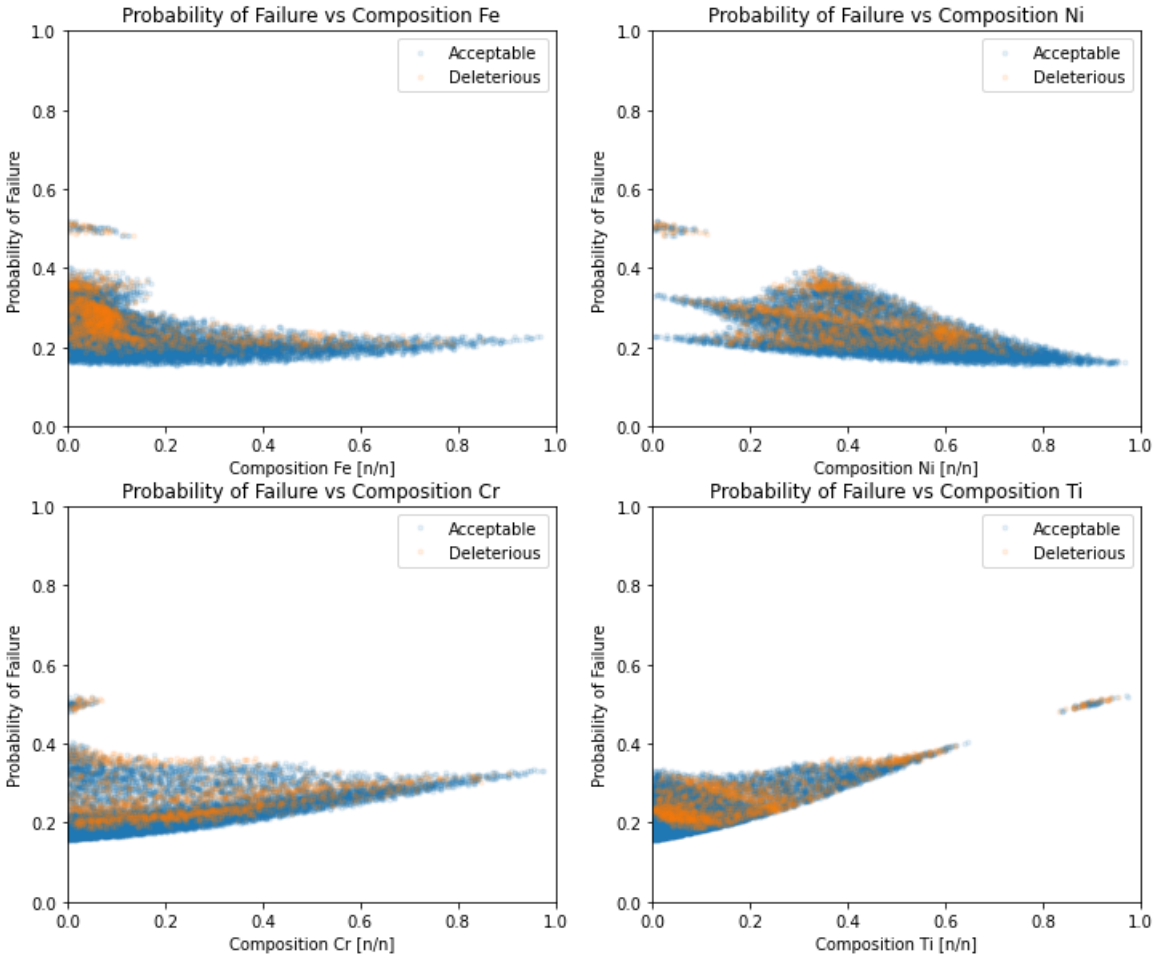

Figure B.3: Filtered Probability Estimations based on the original Gaussian Process Classifier over the entire sample space and any samples that exist below the hard cut-off temperature threshold projected over each material composition. Blue = Acceptable Phase, Orange = Deleterious Phase.

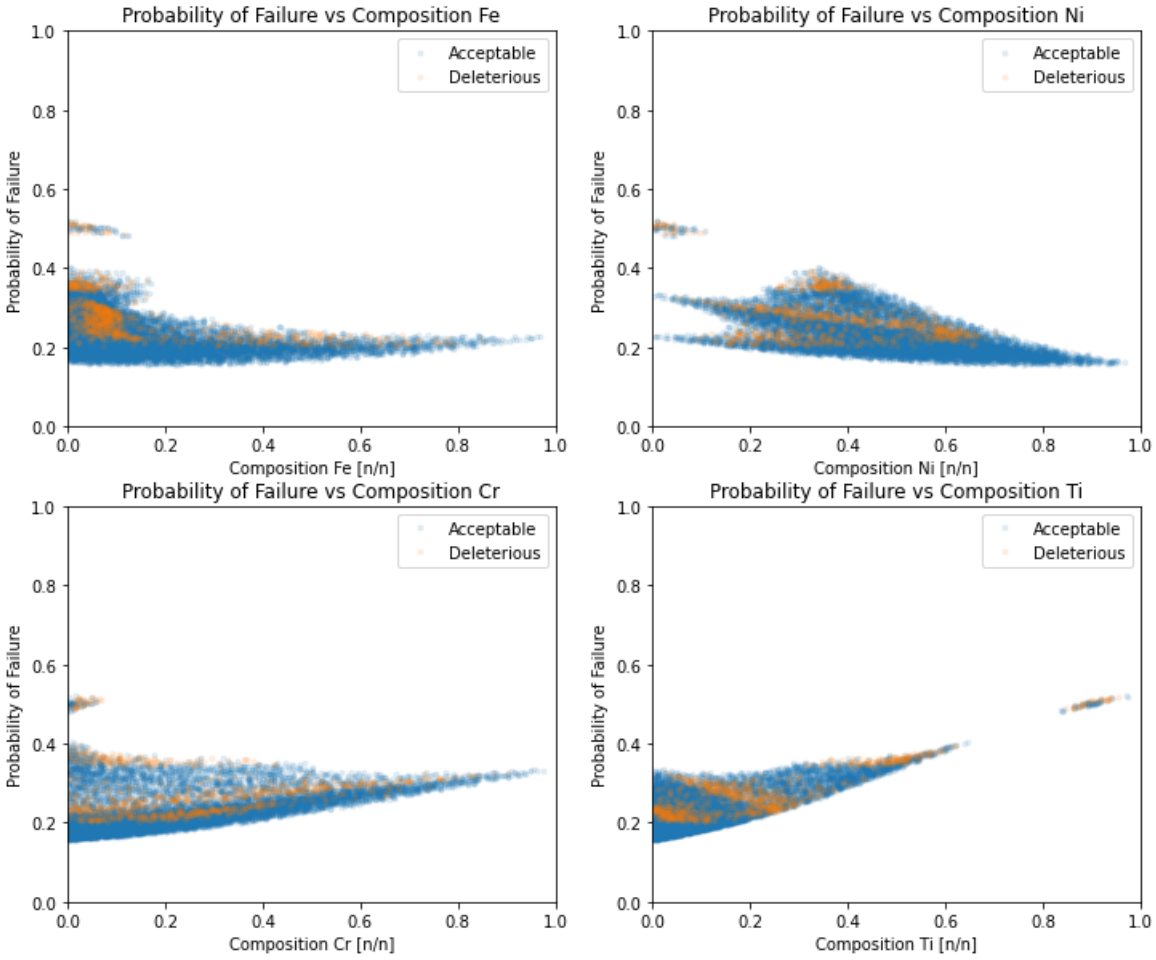

Figure B.4: Filtered Probability Estimations based on the Gaussian Process Classifier over the entire sample space and the second classifier within which temperature is incorporated projected over each material composition. Blue = Acceptable Phase, Orange = Deleterious Phase.