Predicting International Normalised Ratio for Trauma Patients on MIMIC-III Dataset using Machine Learning

Project Report for CHEME 5999

Aasim Ayaz Wani

Net ID: aw579

Contents

List of Figures

List of Tables

1. Introduction

Anti-coagulants are one of the more profitable drugs currently in the pharmaceutical industry with a market cap of 27 Billion Dollars and Warfarin alone covers approximately 73% of the market. They have been used for treating thrombo-embolic events, atrial fibrillation, pulmonary emboli among other things[1]. Warfarin is an effective anticoagulant because it reduces the regeneration of vitamin K from vitamin K epoxide by inhibiting the reductase enzymes in the vitamin K cycle [2], [3]. When a person gets injured resulting in some bleeding, the body initiates the clotting process called homeostasis. Two main components influence this process - platelets and coagulation factors. They both get activated due to a damaged cardiac wall or an injury. Platelets are cell fragments that result from the disintegration of megakaryocytes.[4] Normally the body produces chemical messengers that inhibit platelet activation, but, when injured, the body's internal mechanism releases coagulation factors that direct platelets to the injury site to prevent blood loss. The "Activated Platelets" merge with each other in different shapes to form clusters around these injury sites. This restricts the blood flow out of the wound. To make these clusters more strong, weaker clusters of platelets merge together and in the process reinforce each other by the production of fibrin by the clotting cascade. These get tangled up with the platelets in the plug to create a net that traps even more platelets and cells. Thus because of these, the clot becomes tougher and more durable. Vitamin K is a necessary variable that has a direct impact on the production of clotting factors II, VII, IX and X[2], [3]. Patients taking warfarin are rendered deficient in regenerated vitamin K. So clotting factor synthesis is therefore likely to be critically dependent on the oral intake of the vitamin [2], [3]. But even 70 years after first developing the warfarin drug, the interactions which it has other drugs are not fully understood. Even after this uncertainty, Warfarin remains one of the most commonly prescribed oral coagulants in the market. The reason being that even though warfarin is a very effective drug, it is affected by a whole list of variables some of which are measurable and some of which are immeasurable. To better predict the effect on INR for specific warfarin dose, a lot of studies have been collected. Although Warfarin has been a widely adopted drug because of its effectiveness in reducing risk, how suited a particular warfarin dose is for a person still depends upon its effectiveness in maintaining the INR <2. Warfarin reduces the ability of coagulation factors to interact with the phospholipid membrane[1]. To quantify the extent of the coagulation ability of the body, a prothrombin time (PT) test is used. INR

(International Normalised Ratio) test compares the PT time of the blood of the injured person and compares it with that expected in a normal person. INR is a dimensionless number that compares the coagulation ability of a person by how much it is greater than 1. Based on the value of INR doctors' can make an assessment of how much warfarin should be given to the patient. Usually, an INR value of 1-1.5 is considered to be in the normal range. If the INR is below 1 it means that the blood is taking less then the usual time to clot and is called under-anti coagulation, and it presents the higher clotting risk. An INR value greater than 3 indicates that it is taking more than the usual time to clot which indicates under-coagulation, and means that there is going to be a much higher risk of bleeding. The issues with an accurate prediction of INR stems from its dependence on a diverse set of genetic variations, biochemical parameters, rate of vitamin K metabolism, Dietary practices, the interaction of warfarin with other drugs. Even though measuring the impact of warfarin by measuring the PT time takes only 10-15 seconds, after administering the warfarin, the change in INR is reflected by the coagulation factors and their half-life varies from 7 hours to 6 days. [5]. Also, warfarin needs to be administered for at least 2 days. To measure the effect of a specific dose of warfarin we need to measure the INR value after the half-life period of the coagulation factors starting from the time warfarin dose is administered to the patient [6]. There are observable changes in INR trends as early as 7 hours, however, they are primarily driven by the coagulation factor 7 which has a half-life of 7 hours [7]. These earlier trends tend to be misleading [5] because it takes approximately 4-5 days for all the factors from the day of administering to measure the complete impact of warfarin dose[8]. Because of this significant time gap between the time of administering and actual impact of the drug it is a challenging problem [6].

2. Literature Review

Warfarin is an immensely popular anticoagulant drug with a stronghold on the market, however a 2008 analysts study indicates that manufacturers are losing as much 500 Million Dollars a year because of the reluctance of physicians to administer the drug [9]. The reason being that the warfarin is affected by a whole list of variables some of which are measurable and some which are immeasurable. To better predict the effect on INR for a specific warfarin dose, a lot of studies have been collected. Even though Warfarin has been widely adopted drug because of its effectiveness in reducing risk, how suited a particular warfarin dose is for a person still

depends upon its effectiveness in maintaining the INR < 2 there is still a lot of uncertainty in understanding what could be the right dose[10]. Considering all this, the doctors when trying to decide on the dose, are always making an educated guess about the dosage and have to do multiple iterations for getting the correct dose. This turns out to be an expensive process because warfarin dose needs to be administered for at least 2 days to check its efficacy and the patient may still be in pain during this trial and error since the clotting mechanism in his body is not functioning properly [10], [6]. Though feature selection to find an optimal number of features that maximize certain criteria has not been used in the domain of warfarin. But considering the reasons which make it a multi-variable problem, it is clear that there is a need for such an approach. [11] worked on the problem of determining an optimal number of features by binding it with an acceptable range of error. They developed a dual objective function problem both of which were driving for opposite objectives. One objective function was trying to minimize the number of variables we are considering for the purpose of the domain, another objective was minimizing the error which we were getting when we considered a particular subset of variables [11]. To confirm the results they used a prior understanding of the domain space and used that to infer their results and compared their performance with the other relevant methods.[12] conducted research on 553 patients they did feature selection using the non-dominated sorting genetic algorithm and multi-objective particle swarm optimization and they evaluated the results using deep neural networks and used mean absolute percentage error, root means square error of 0.109, 0.1 respectively.

The problem with not using these feature selection methods to select optimal features is that there are a lot of possible unexplored variables to consider in a very short period. This is effective in most cases even in real life because Warfarin dosage has been treated as a trial and error based method because this is dependent on a whole list of variables and this is not difficult to do if the patient coming in will have the time[12]. The normal INR value of <3 is also uncertain and this not generally accepted everywhere. Every city has its own criterion and protocols in order to deal with these INR values [13], [14]. A lot of research has been conducted on trying to link a person's diet to the effect on how they interact with warfarin. But still, there is no consensus on the clear impact. Multiple independent previous pieces of research have been contradictory. Some studies found out that there was a positive correlation between maintaining an INR level and vitamin K intake[15]. However other studies indicated

that there was no clear impact but the INR just needed to be above a certain threshold level. The variation this intake difference is linked to the prominence in gene variation [16]. Not much research has been done to identify this variation in the connection of vitamin K response to the polymorphisms in the CYP2C9 AND VKORC1 genes[17]. But a lot of research has been conducted which has been done which shows independently what is the impact of this polymorphism of these genes and linking them to INR values and warfarin dosage [18], [19], [17]. [20] conducted a study on 32 children with ages ranging from 3 months to 17 years. They used mean absolute percentage error and root mean square error. Research by [21] conducted research by giving anti-coagulant warfarin for 24 hours but research by [10] indicates that the warfarin dose should be administered for at least 2 days. [22] conducted research on 50 patients one of both men and women, with an average range of 75 years. Following the oral warfarin for 6 months, their warfarin was reduced to an INR within the range of [2,3]. They had been following the recommendations by the WHO. How different INR has been widely studied because it gets affected the particular gene that a person possesses. This research has concluded gets affected by where the people are from. [23] conducted research highlighting they could control the INR value < 5. They had assumed that if the patients have an INR <5 this would substantially reduce the risk of clot formation. They wanted to see if they just instead of just giving Warfarin dose in 1 installment divided them smaller installments by making smaller adjustments to make sure so that value continues to remain <5. They divided the dose into two types of inputs and this shift resulted in the percentage of people > 5 INR from 23% to just 9%. The objective here was not to get a fixed value of warfarin dose of INR but to prove that continuous input measurement had better performance in keeping the patients within the INR 5 range for more time.

One of the bigger issues which had stifled a better understanding of the domain is that there exist multiple patients coming in with different patient histories it is not possible to make a comprehensive enough model that can take into consideration all these patients simultaneously. For example research by [24] conducted research on the effect of smoking on a patient with atrial fibrillation. They had 117 nonsmokers, 23 light smokers, and 34 heavy smokers they had expected that a heavy or medium-level smoker would have a higher warfarin dosage requirements [25]. The results where inconclusive since all three groups of smokers had some people who had high, low, medium-dose requirements. Looking at this problem previous researchers have tried

to make multiple patients models. [26] conducted a study on a cohort of people ranging from an age of 19 years to 60 years. They initially gave them a certain amount of warfarin for a week during which they continued to smoke but following that week they stopped smoking and then the same amount of warfarin was administered, they found that the INR value was much more stable and reached steady state much faster without the need to increase the dosage.

Another example is the effect that age and gender have had on predicting the effect of INR values. Previous research showed that both age and sex play a significant role in determining the appropriate warfarin dosage. However, the relationship relating to these variables is non-linear. Research by [27] indicated that base on the type of gender they divided the age into different smaller batches and they fitted a linear model to check feature importance [27]. The dataset had been 53% men and 47% women. These multiple patient models basically initially use some heuristic to separate patients then different models for them. This has also been a general cause of concern of researchers since they were for a long time not able to transfer the learnings from one patient and somehow incorporate them into the leanings of other patients as well. Most patients who are given warfarin coming in the emergency room do not have the time to be tested for other drug interactions. But Warfarin has very complicated interactions with other drugs, in some cases, it could be as simple as just needing more or less warfarin dose based on the type of interactions or sometimes it could be potentially life-threatening. [28] conducted a 5-year study on studying the interaction of Ciproflaxin with Warfarin. They found out that people taking Ciproflaxin for gastroenteritis would have a much more prolonged prothrombin time and these patients were all such people who had been previously having normal prothrombin time.

3. About the Dataset

The MIMIC-3 version 1.4 database was used as a source of data. Composed of 28 tables, MIMIC-3 has data for 61,532 intensive care unit patients. It provides information about patient's admission in-out log entries, lab values, demographic information, doctor's notes, how treatment progressed. The patient data is sequential for each unique patient and not the entire dataset. On reading the data from the tables we have multiple entries for each patient at multiple time stamps. MIMIC has a lot of missing values for the patient's test values.

The most likely reason for missing values in the MIMIC data is that patients are coming

in with different levels of trauma coming into the hospital or they had undergone the tests but the records were not entered. Based on the effect of their injuries the doctor recommends order them to undergo a series of tests. Patient data for every unique patient is a time series problem but when averaged we get a single value for each tribute of the patient. This eliminates the notion of sequential nature from our problem. Thus the problem definition can be transformed from a time series problem into a regression problem. The latter case can be viewed as a watered-down version of the time series problem. We are pursuing this route because there are more intuitive machine learning algorithms that can help us in understanding our results since we are concerned about the validity of our data. However, the time series case gives us far more accurate results since we are using prior knowledge about the similarity between these data points.

4. Merging Approaches

The way the database is structured enables the user to extract the data about the patient details and one variable at a time. The MIMIC database table from which we have extracted has data for all the patient biochemical parameters entirely in 1 column labeled "Value". To query the values of that particular lab test we had to query the data by filtering for a specific value of another column "Item Id" which stored unique values for every lab test done. With that information, we also tagged the subject ID and in some cases, I have also queried the chart time so that we can maintain order in which data was entered. There are 3 types of merging approaches which we have pursued:-

1. Outer Merging on Primary Coagulation factors
2. Inner Merging on Primary Coagulation factors and Biochemical Factors
3. Outer Merging of some Coagulation and Biochemical Factors

4.1. Outer Merging on Primary Coagulation factors

4.1.1. Database Preprocessing

In the process of "outer merging" the data we don't lose any information from the data, this is because we are outer merging on both "Subject id" and then "Chart time". Using this method we can combine such data entries in which both these happen to be the same. Thus we are not

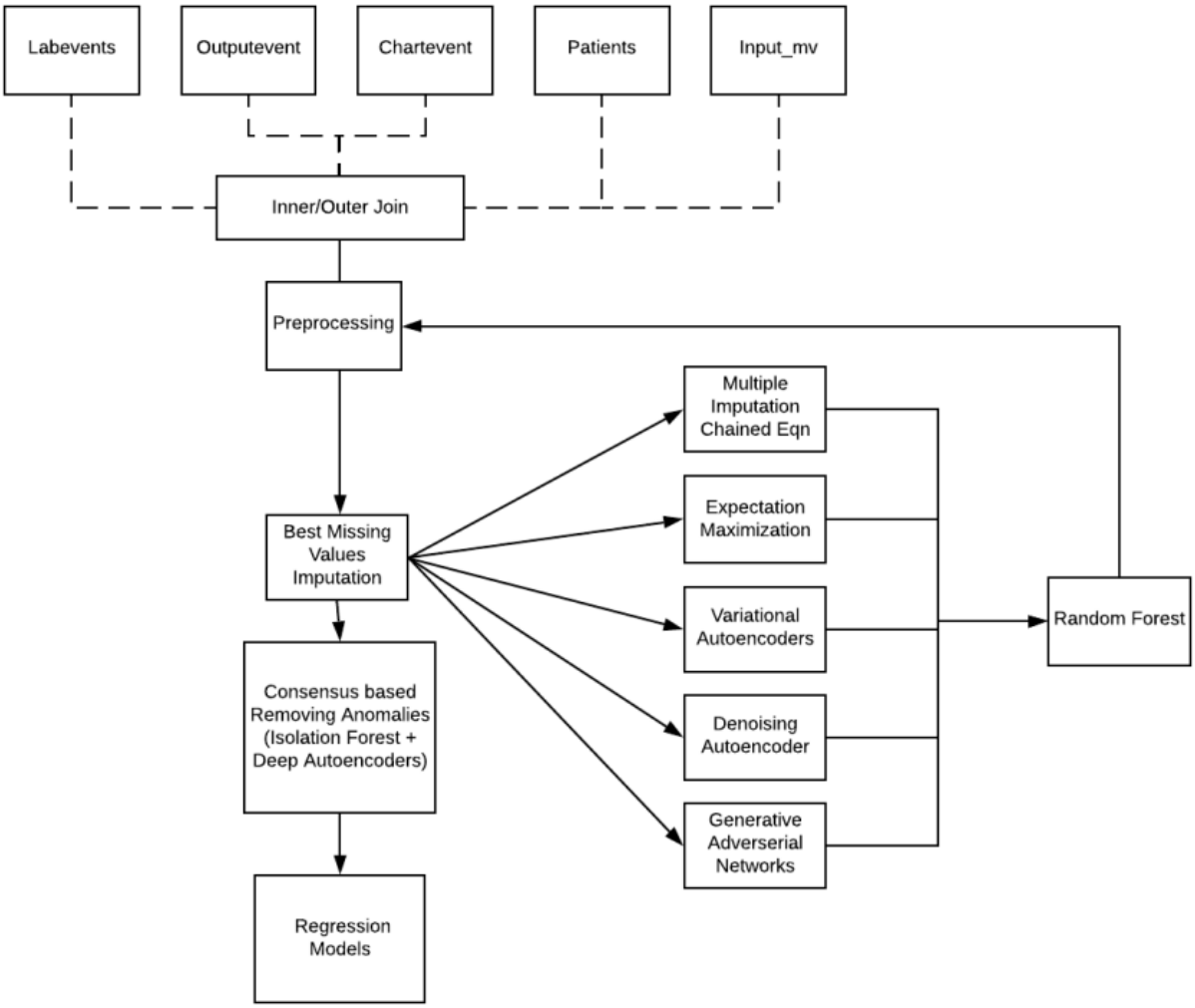

Figure 1: Flowchart of Process Methodology

losing the time-series nature of the data such that for every patient there exist multiple entries of the patient. The current size of the data for this database is 475,025 rows and 16 columns (this includes only the coagulation factors).

4.1.2. Discussion

Looking at the above plots figure 1 and figure 2 we can see that the dataset has close to 98% of the INR data, and subsequently, the next few most populated variables are factor 7, Anti-thrombin and Thrombin they only have close to 1% of the data. The other variables are having only $< 1\%$ of the data. Considering how much of the data is missing any standard imputation cannot be used since such an extent of the data is missing. Because if we use any standard imputation it will cause the model to lose any information since the variation in

the dataset will get lost. So, it is not that unexpected that the standard imputation is one of the worst-performing methods. Considering that I have used multiple imputation methods mainly because they can take into account the relation in between variables when filling missing values which is in contrast to standard imputation because they assume independence between variables. Since we have the raw data with variables on different scales we cannot have them

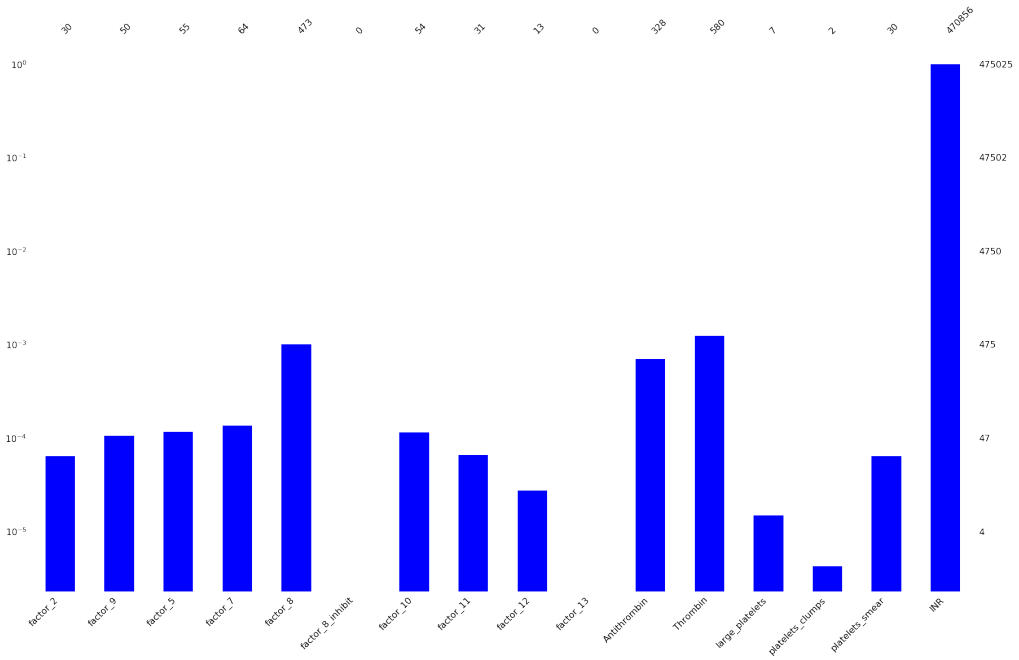

Figure 2: Analysis of Stage 1 Training Data

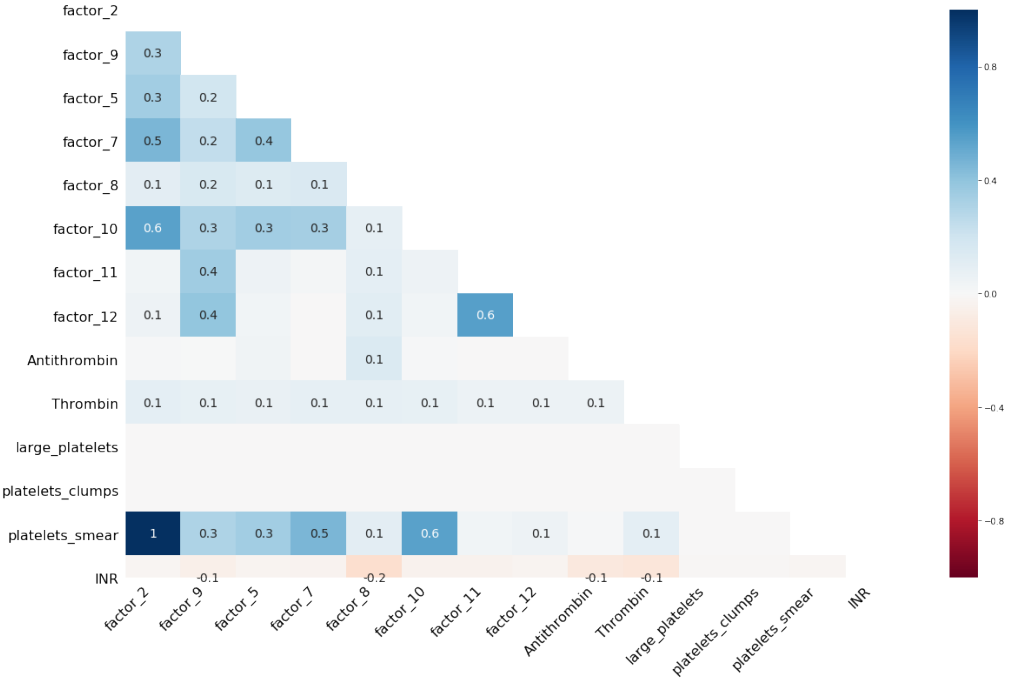

Figure 3: Heat Map of Stage 1 Training Data

directly going through the principal component analysis and outlier detection methods. This is because these methods tend to be extremely sensitive to outliers and get affected by the presence of uneven scaling. This is why we have both performed min-max scaling and performed normalization. The above plot is a plot showing Principle Component Analysis (PCA) with 2 components, the objective of doing that is to reduce all the variance in the data and remove the noise. PCA is a popular dimensionality reduction method it reduces the number of variables by projecting the data from n dimensions from higher to lower dimensions. In most cases running only 2 components case is sufficient but we ran the test to check how much of the variation is explained by the 2 components. In our case, the variation explained by only 2 components is close to 100%.

After that, I used the anomaly removal method called "Isolation Forest" which can be used to remove points. This is a powerful method because we have a lot of data and we don't want to spend much computational power on this because we already have 4 $\times 10^6$ rows and this method doesn't have to train on the non-outlier or outlier points using this knowledge we can identify points. But this method eliminates the cost of finding any distance function over the entire dataset making this a n^2 problem which will take many iterations and will be expensive. The isolation forest with 0.1 contamination reduced the size of the dataset by slightly less than 9% of the data points were removed flagged as outliers. There has been a lot of previous research done on performing multivariable linear regression on just the coagulation parameters. In this part of the study we considered multiple coagulation factors (2, 5, 7, 8, 9, 10, 11, 12), Antithrombin, Thrombin, Large platelets, platelets clumps, platelets smear. Since the p-value of all the factors being mentioned has a value of less than 0.005 which implies they are all significant. Looking at just the coefficient for the variables we can see that platelets smear, factor 7, factor 2 is the most significant positive variable and factors 2, 9, seemed to be the most significant negatively effecting the INR value. These results are deemed to be most consistent with previous work done.

4.2. Outer Merging of Coagulation Factors and Non-Coagulation Factors

4.2.1. Database Formation

In this process, we are using outer merging on different columns of data. When we are "outer merging" on the data using coagulation and non-coagulation parameters, we don't lose any

data. This is because any instance in which the subject id for the two entries match, then a new entry will get added to the new database either associated with some specific time entry if that also matches, in case it does not match then with the same subject id a new row gets added but with the time entry for existing time.

4.2.2. Discussion

In this data, we have been able to preserve the time-series nature of the data. Time series data class of temporal data arrangement in which observations are being entered chronologically. This presents additional information about the data by the sequence in which they appear [29]. But when there are multiple sources in the dataset in sets of ordered data this does not remain a standard time series problem since they are mostly made for single-source datasets. A shorthand approach to solving this problem is to restating this situation as a regression problem by using a set of transformations that introduce a new feature in the data which provides an index of the position of that particular element in that source. This work has already been reviewed and applied by [30], [31] in both this reformulation approach and the direct simplification approach for identifying these trend points in data to identify these sources end and starting points.

In this linear regression, we have used both the coagulation parameters and the biochemical parameters to predict for INR values. Very surprisingly I found that p-value for most of the parameters has a p-value of greater than 0.005. The positive coefficient factors which affect INR include factor 2, 5, 8, Anti-thrombin, large platelets, bilirubin, Oxygen, white blood cells, white blood cells count, Potassium Data, Potassium blood percentage.

4.3. Inner Merging over some Coagulation Factors and Non-Coagulation Factors

4.3.1. Database Formation

In this process, we are using outer merging on different columns of data. When we are "outer merging" on the data using coagulation and non-coagulation parameters, we don't lose any data. However, the Patient data, Lab data, for every unique patient is a time series problem but when averaged we get a single value for each tribute of the patient. This eliminates the notion of sequential nature from our problem. Thus the problem definition can be transformed from a time series problem into a regression problem.

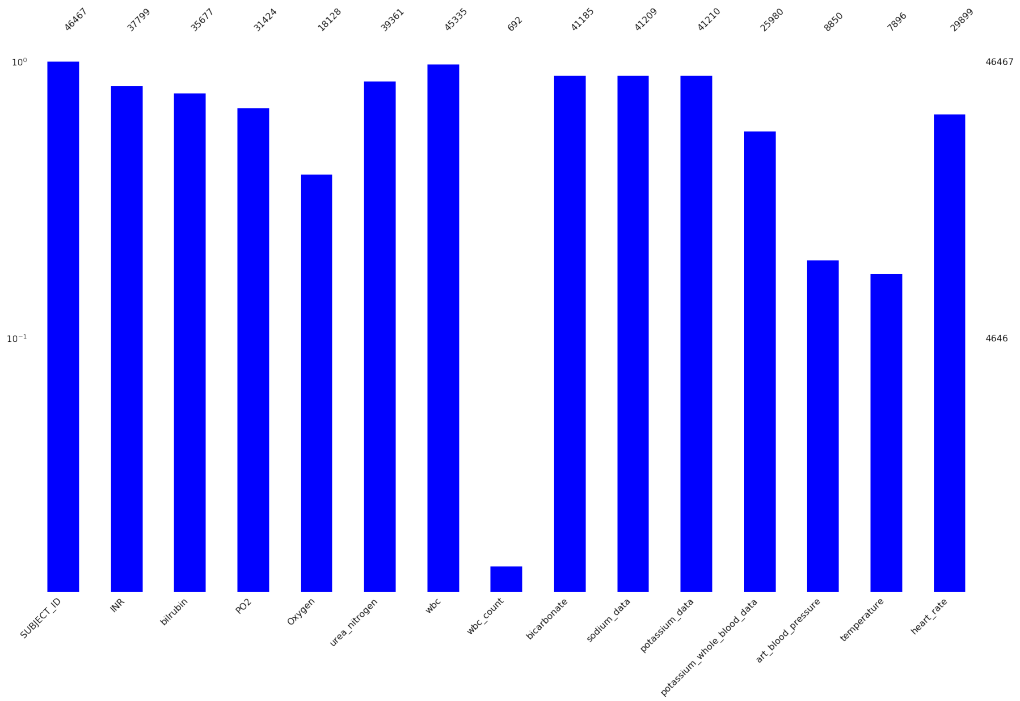

Figure 4: Analysis of Stage 3 Training Data

4.3.2. Discussion

From the theory, we do know that the INR can be divided into 2 parts, people within the feasible range of $<$ or $>$ 2-3 INR. Even though there is no clear classification at INR 3 but we can ascertain from previous research that the people below $<$ 3 INR are generally considered to be safe. To confirm our hypothesis we are using the elbow plot which also showed that the 2 is the optimal number of clusters. Even though the elbow plot is not a confirmatory test but this test was taken with prior knowledge about the data we can be sure that the Looking at the Principal Component Analysis of the data and setting the number of clusters at 2. Since we already know that PCA assumes linear assumptions about the data. From the above plot, we can see that the data seems to distributed as non-linear data but we can see that from the PCA of the normalized data, I wanted to perform TSNE on the data to actually visualize the data.

5. Methods of Feature Reduction

There are mainly 3 feature reduction methods which I have used - Principle Component Analysis, Autoencoders and TSNE (T-Distributed Stochastic Neighbour Estimation). Dimensionality reduction techniques are used primarily for two reasons - visualizing data when problem space

is high dimensional and to retain only the most essential aspects of the data.

5.1. Principal Component Analysis

Principle Component Analysis is a statistical technique that seeks to generate such a linear combination of data to maximize the variance and minimize the reconstruction error. This is achieved by trying to minimize the reconstruction error, we are automatically maximizing the variance in the data. PCA is commonly used either for the purpose of feature reduction by finding such an axis in which maximum variance in the data is preserved or for finding potential outliers. Even though we may go to as many dimensions as possible but in most cases, we only tend to take as many dimensions till we can reach $> 95\%$ of the % variance explained. The main concept is that PCA can isolate a small number of common components that can condense the variance information of the data this is especially true when analyzing large datasets[32].

5.2. t-SNE

t-distributed stochastic neighbor estimation abbreviated as t-SNE is a dimensionality reduction method that is preferred for data visualization of high dimensional data. Essentially t-SNE works similar to PCA but instead of using linear relation among variables it tries to fit joint distribution among variables and tries to reduce the kullback Liebler Divergence among the joint probabilities. However, it usually takes many iterations to find an optimal value of t-SNE, also since this is a non-convex problem we will get different values on different initialization points. This is in contrast to PCA where we don't need to find any optimal hyper-parameter like perplexity and the number of steps. Considering the way t-SNE works it is not used when the dataset size is too high. I have used tsne as the last step resort for feature reduction when PCA fails to create feature reduction. Even though tsne has many hyperparameters to tune I have only been using perplexity. Perplexity is the measure of information which is 2 raised to the power of Shannon entropy. My objective was to separate the data into two distinct clusters, however, if perplexity is too high represents data in form of equidistant structures(points, etc), if too high it might transform into one single spherical ball.

5.3. Autoencoders

Simple Autoencoders are also a variety of dimensionality reduction methods. In which data after being passed through a certain number of hidden units in a layers gets successively passed through an increasingly smaller number of hidden units and after few hidden layers its gets

projected back into the original dimension from where we had started. This process of initially reducing the number of dimensions and then projecting the same into higher dimensions even though looks fairly simple has the benefit of reducing the unnecessary sample information and keeping only the very important aspects present in the data. The first stage is referred to as encoders and the second process is called a decoder. This reduces any added noise by maximizes the reconstruction likelihood and minimizing the reconstruction error but the innovation in filling the missing values came in this field very recently. Autoencoders can be of mainly 2 types - under Autoencoders or shallow learners and over Autoencoders or Deep Autoencoders. The latter has more hidden units than the number of features, and prior has less number of hidden units

This is because there is no structure in the latent space there is no restriction on the latent space to maintain the continuity or even allow easy interpretation. Because of these similarities, the variational autoencoders are also compared with PCA one important difference between them is that in contrast to PCA there is no constraint on the components of Autoencoders to be orthogonal to each other. In variational autoencoders we can learn the smooth latent space representations of our original data, this ensures that at every point in the latent space there is a probability associate that relates to something in data. This guarantees that we can every point in the latent space does represent some of our initial data.

If we just try to minimize the reconstruction error by trying the KL divergence term between the prior and the latent space to have similar distribution we would just be copying the initial data and we would have failed to understand the underlying distribution. Variational Autoencoders try to act in the middle of these two extremes by trying to drive the KL divergence to be as small as possible while trying to generate a smooth latent distribution for the data. When we try to optimize both of these constraints we are essentially keeping KL divergence as a regularisation force.

We have used feature reduction methods not only to visualize we thought that since we have such an enormous number of missing values we want to remove all the unnecessary information from the database. We are using this especially after the process of filling missing values because of the extent of the missing values we are very concerned that the data may actually just have been overfitting on the data and the model does not learn enough from the data. So to reduce the chances of overfitting and still retaining some of the data we are using other dimensionality

reduction methods and I am relying on feature reduction method to retain only the essence and remove all the unnecessary similar repetition in data.

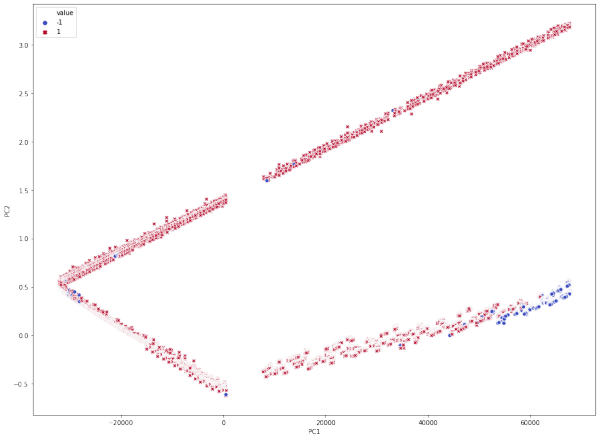

Figure 5: Outlier Detection on Stage 1

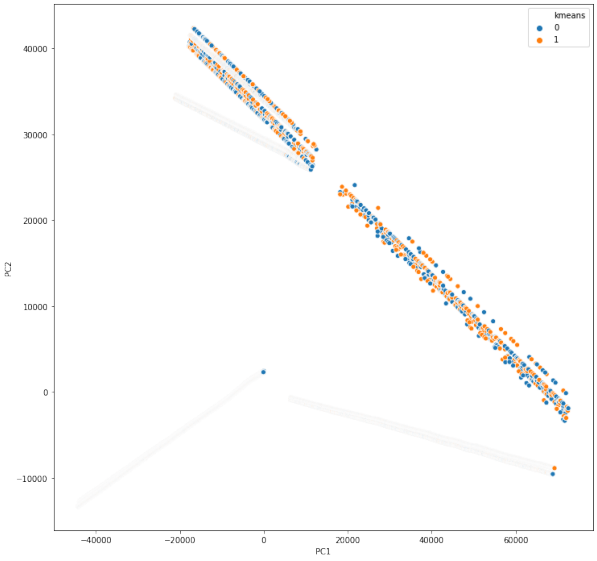

Figure 6: K-Means on Stage 2

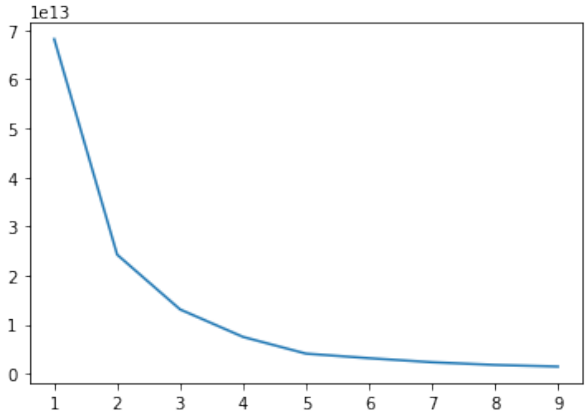

Figure 7: K-Means Elbow Method

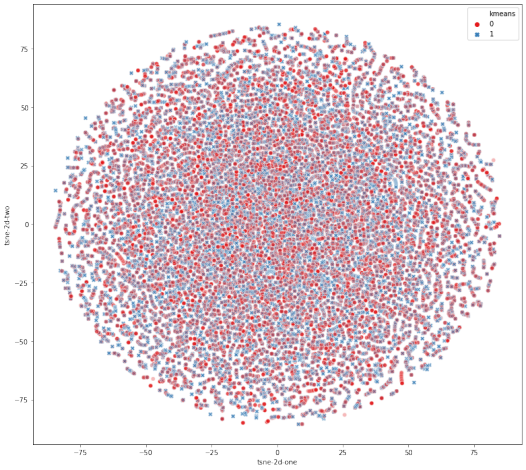

Figure 8: t-SNE with Perplexity 5 Stage 2

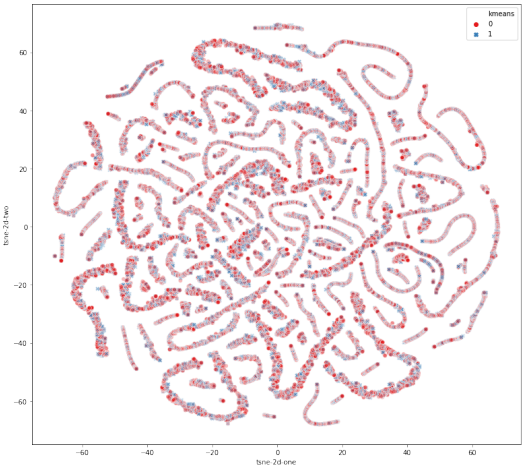

Figure 9: t-SNE with Perplexity 50 Stage 2

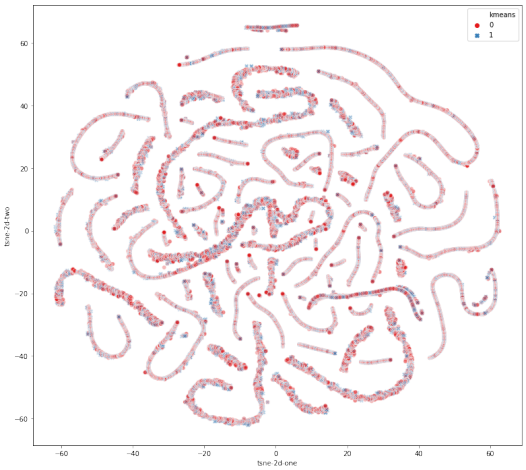

Figure 10: t-SNE with Perplexity 100 Stage 2

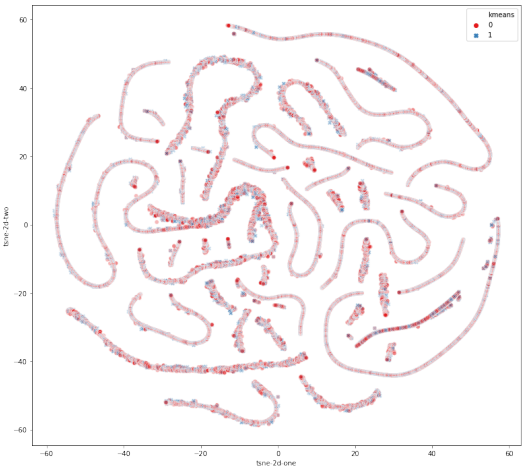

Figure 11: t-SNE with Perplexity 150 Stage 2

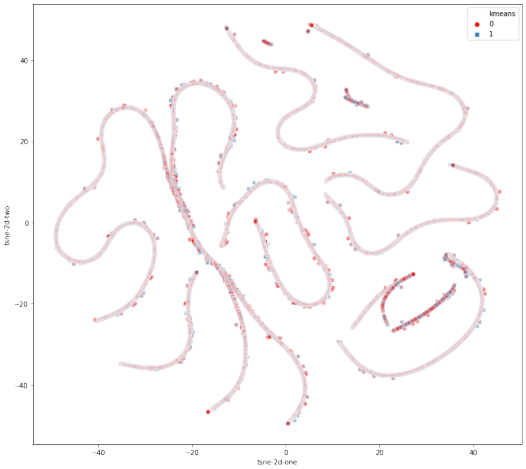

Figure 12: t-SNE with Perplexity 400 Stage 2

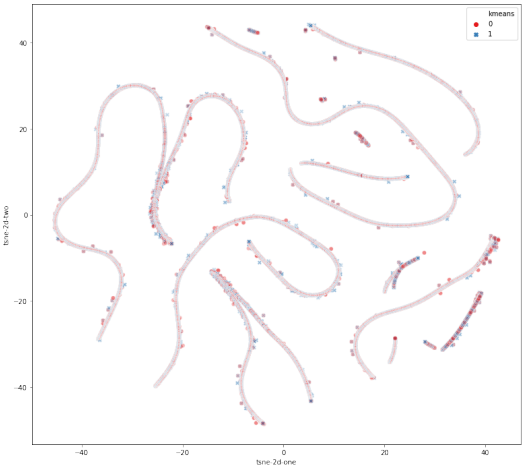

Figure 13: t-SNE with Perplexity 500 Stage 2

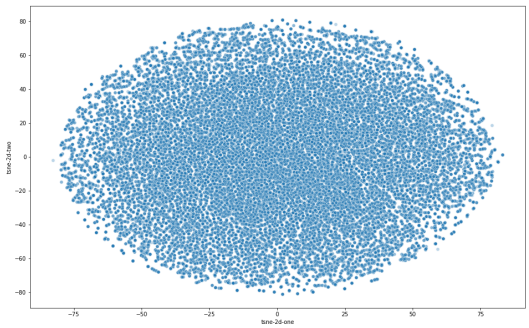

Figure 14: t-SNE with perplexity 5 Stage 3

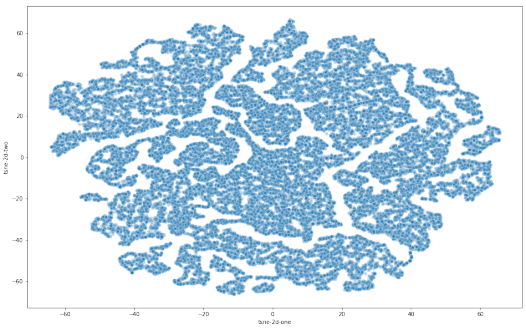

Figure 15: t-SNE with Perplexity 50 Stage 3

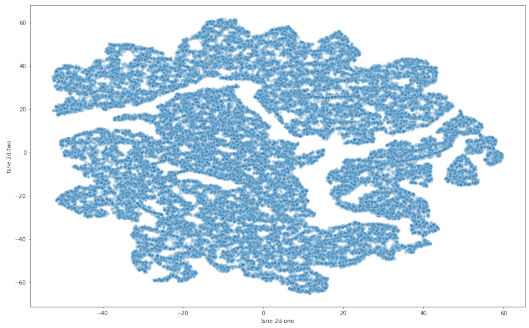

Figure 16: t-SNE with Perplexity 100 Stage 3

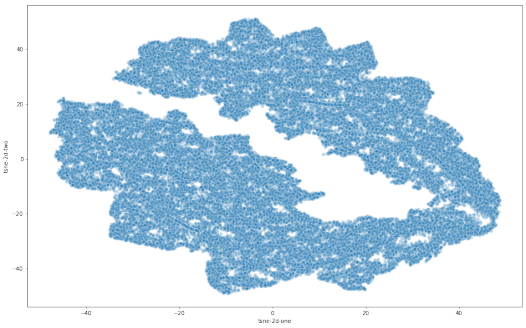

Figure 17: t-SNE with Perplexity 260 Stage 3

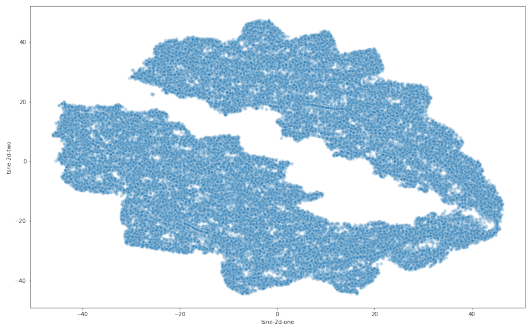

Figure 18: t-SNE with Perplexity 350 Stage 3

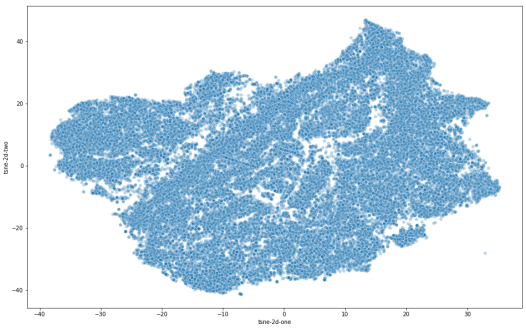

Figure 19: t-SNE with Perplexity 400 Stage 3

6. Dealing with Outliers

For outlier detection, I have used a consensus method of Isolation Forest and Deep Autoencoders. Isolation Forest generates outliers by splitting data by randomly choosing a split value between the maximum and the minimum value of features. The quality of the split depends upon how much time it would take to separate the points clearly. If a normal point is misclassified as an outlier it would be much easier for the model to understand the rule for the model, on the other hand, if a point is indeed an outlier and is very much on the edge when it comes

to feature values it would not be very difficult for the model to generate rules for the same. For Deep Autoencoders we have 6 dense layers. How it generates outliers is by dividing the data into smaller batches and measuring if the data has more error then the average accepted error it gets tagged as an outlier otherwise passed. However, the whole batch gets classified as an outlier. To fix this problem, I added a consensus. The need for outlier detection stems from the fact that in a complex system like predicting INR or monitoring warfarin dosage this is bound to have a lot of outliers either caused due to noise or since we are dealing with medical data sometimes values are just entered wrong. But there exist so many non-linear and linear relations that it is very difficult to make any assumption about the underlying relationship so we are using isolation forest which makes minimal assumptions. Also since we are having a large dataset we want to use a method that is more computationally cheaper to run.

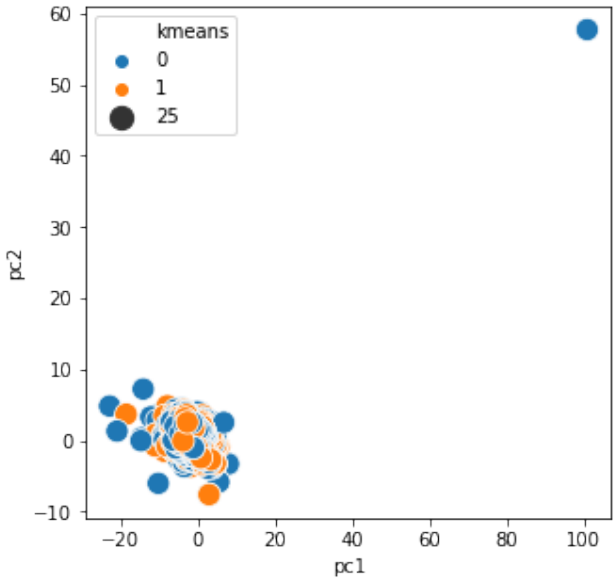

Figure 20: PCA on Stage 3 Data with K-Means

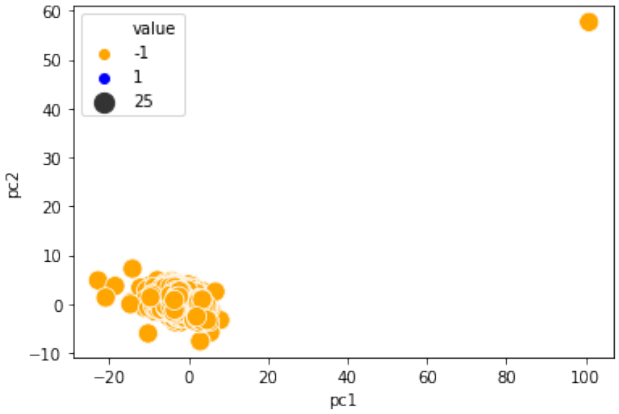

Figure 21: Visualising Outliers on Stage 3 with PCA

On applying Principal Component Analysis on data we wanted to see how well does the data fit when we are trying to use a linear method to fit them. From the prior understanding of the domain space, we already know that there should be two clusters on trying to visualize this we found this is not a good representation for the data. Then I realized that maybe this could have been due to the presence of outliers in data. So I used Deep Autoencoders and Isolation Forest consensus method again to weed out the assumed outliers. To see which points were being removed by this consensus, I again resorted to using PCA for the visualization.

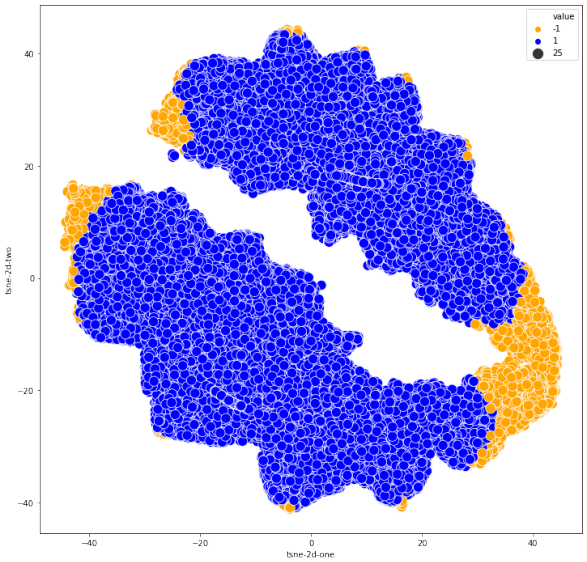

Figure 22: Visualising Outlier Detection on t-SNE

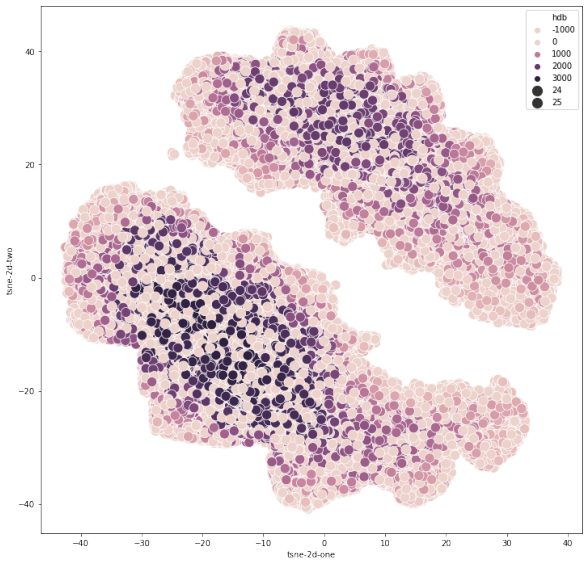

Figure 23: Hierarchial Density Clustering on t-SNE

Figure 22 shows the results which I got, from this figure we can infer that this visualization and largely this outlier detection method has been a failure. The outliers should have been following a distinct method that should be separable if this particular feature reduction method would have been a good fit for the data. But since this is not visible we are moving to other non-linear dimensionality reduction methods like TSNE.

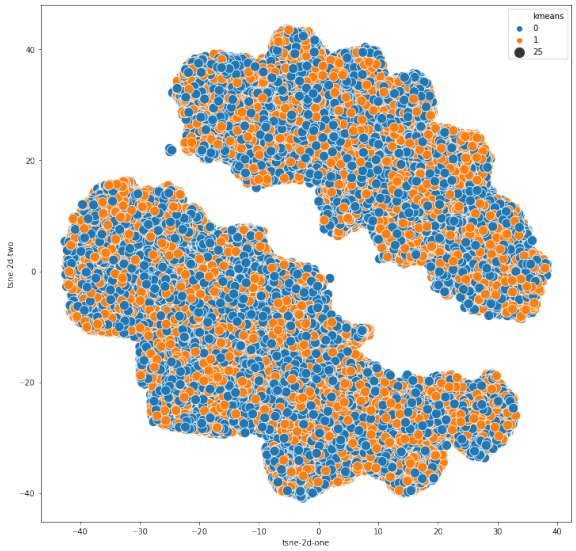

Figure 24: K-Means Clustering on the data after t-SNE

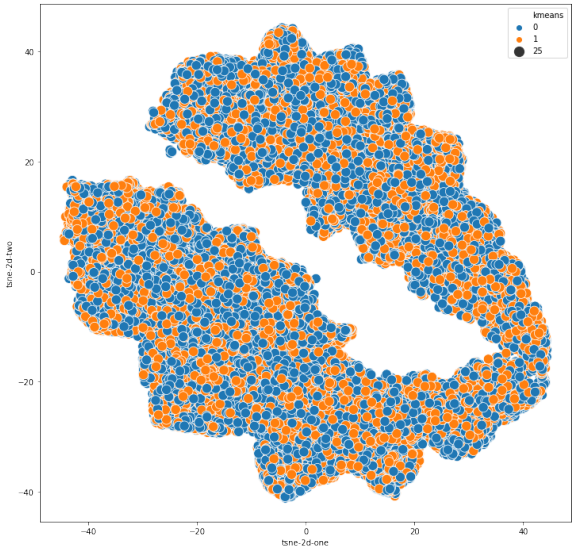

Figure 25: K-Means Clustering on the data before t-SNE

7. Dealing with Missing Values

There are mainly three types of missing values - missing completely at random (MCAR), missing at random (MAR), Not missing at random (NMAR) [33][34]. MCAR assumes that the missing values of a particular variable are not affected by the values in other variables either observed or unobserved. Instead, these are assumed to be generated through a process such that their presence or absence will not have bearing on increasing the bias on the model. Also, since the probability of missingness is the same for all missing data points. MCAR has a very strong modeling assumption which is difficult to prove as you would require access to the missing values to ascertain their effect [33]. For example, the patient goes in for a consult he goes to get some blood tests conducted on him, but instead due to some malfunctioning equipment, his test results show only partial results. MAR assumes that the missingness depends only on the observed variables, This means that values of missing data can be entirely explained by other observed variables [35], [33]. MAR assumes there is a systematic relationship between the propensity of missing values and the observed data, but there are no underlying reasons for the occurrence of the missing data. MCAR and MAR are both considered 'ignorable' because we don't have to include any information about the missing data itself when we deal with the missing data [36]. In the case of MNAR, the values of these missing data points depend on both observed variables and the unobserved variables. MNAR is also called a "non-ignorable case" because the missing data mechanism itself has to be modeled as you deal with the missing data. For example. a sensor may not be picking up readings during some time in winters, assuming that this is MNAR means that the very fact that values are missing in winters is itself a variable. [37], [38], [39], [33], [40], [41], [36].

There are 2 types of modeling approaches - Discriminative and Generative models. Discriminative learning models learn from the mapping of the feature space from Y→X [42]. Such as Logistic Regression, Support Vector Machine, Neural Networks, K-Nearest Neighbour. On the other hand, Generative learning is such algorithms those which learning the mapping from X→Y. Such as Naive Bayes, Linear Discriminant Analysis, Bayesian Networks, Hidden Markov's Models [42]. In general, generative models are considered to be more appropriate under such circumstances when we are trying to identify a relationship between variables and the order in which order has been added to some prior understanding of the missing values this facilitates filling the missing values. There are many ways of approaching missing values, in

most cases imputation is needed, however sometimes missing values may be deleted altogether. The appropriate choice depends on the identifying the process which has led to the creation of the missing values an how much does ignoring these values could impact variable relationship i.e does it introduce or remove information in the data which may deviate from the true data.[33]

7.1. Dealing with Missing Data

Dealing with missing values can be a challenge in time series data since when we are using any imputations like standard or multiple imputations are making an assumption about the data being from the source. In the case of our particular domain what it means is that we are assuming that every patient we which did not get this test ordered by the doctor all had the same reason for doing so and in doing so this could lead to problems. Since the data which we have for every patient is already averaged data [33]. In normal time-series data, we are having one source and we have sequential data for that, this means that the normal time series data cannot be applied because they assume that every entry is having the same source since that is not the case when multiple sources are involved. The brute force solution to solving this problem would be to run a time series data separately and make an ensemble model of them. But this would be an exponential growth algorithm. However there are still ways of retaining the time-series nature of the dataset by artificially inserting a new column into the data which just reflects the data entries position in the dataset, this process is called time sequence reformulation [30].

To better be able to assess the working of our approach we are converting the regression model into a classification problem. This is because in a regression model we cannot get the exact performance metric it is the more relative method which doesn't provide us information about how this method To assess the working efficiency of different approaches for filling missing values I needed to used the same approach for every dataset for validating the results. For this purpose, I used random forest with 100 estimators and max depth 5. We used 20% of the data to test the data.

There are two ways of filling out missing values either we could use the model embedding from original data or we could use the averaged data. One possible reason when averaging could be useful would be using when we are having If we choose to go with the prior one we have the advantage of using the raw dataset. We are assuming that each patient The average outpatient dataset has the advantage that we don't have to deal with the entire dataset at the

same time since on averaging we are making the assumption that the reason why every patient came to the hospital is considered to be a distinct value independent of the other.

Methods such as Denoising AutoEncoder, Variational Autoencoders, Generative Adversarial Networks all rely on their not being any missing values in the data. The way they are helpful is that looking at the relationship in the remaining data they try to leverage a relationship that might be more closer to true data which is not known to us. The implementations which we have used for the above-mentioned algorithms did not specify the usage of any specific method, the original implementations of the DA and GAN actually rely on mean but since I have less 5% of the data in some variables I felt that it would be better to rely on using MICE since it would at least try to model the effect the other variables which standard imputations did not ensure.

7.1.1. Deletion Methods

In some cases instead of retaining the entire dataset, we are more concerned about the quality of the data. In such cases we use *Case wise deletion* (CWD) methods, these deal with removing units of data who have some information missing. There are two types of CWD – *Listwise Deletion*(LWD) and *Pairwise Deletion*(PWD). In former we discard such every entry for a row in which there will be at least one-row value missing, in latter such data points are deleted which contain missing data, for the analysis being carried out. CWD uses MCAR assumption is very difficult to prove as a consequence the analysis usually produces very biased results less generalizable models [43], [44], [38]. Especially, if the missing data are existing concurrently this could severely limit the model performance and would severely limit the performance on unseen results. Among these two, PWD is preferred since LWD can lead to loss of considerable information if the number of missing values is high. Since the missing data is eliminated completely it not creates not only increased biased but there is also uncertainty about how much has the quality been affected which adversely effects models based on the final data [5].

7.1.2. Single Imputation

In Single Imputation (SI) we are assuming that every missing value is having the same value; which is calculated by using some statistical methods like mean, median, mode. A major flaw in SI is that the same values are imputed regardless of the variation in other variables because of this SI model does not take into account the uncertainty because of which it increases the

bias in the model. This is because we are tricking the model by giving it more data than we have since we don't add any new information into the dataset. This is especially relevant when the number of missing values is a significant portion of the dataset. Single value substitutions assume that the data is MCAR but it could lead to unreliable estimates about the variance of the dataset and correlation between variables [41], [45], [46], [47], [38], [43].

7.1.3. Multiple Imputation

A major flaw in single imputation is that it assumes that every missing value is given only one value based on some statistical method. Subsequently, these missing values are given the same consideration as non-imputed values [43]. This is overcome by replacing missing values with multiple imputed values, reflecting our uncertainty about the imputation process until the error is below our accepted threshold and further iterations don't improve the data quality. This approach is called multiple imputation[41] [45], [39],[45],[43]. **Multivariate Imputation by Chained Equation** (MICE) generates different versions of the same dataset at each step of iteration [48]. Imputation is done by drawing from the conditional distribution of the variable with missing values(VM), with every other remaining variable. Imputation over the missing variable is done by calculating values from the posterior predictive distribution over VM[48],[45]. The process is repeated for each variable with missing values in turn: one such round is called a cycle. Because each variable is imputed using its own imputation model, MICE can handle different variable types (for example, continuous, binary, unordered categorical, ordered categorical)[48]. **Expectation Maximization** is a Bayesian method which seeks to maximize the likelihood function for the distributional parameters of some multi-modal data. It provides a point estimate of the parameters and assume that the missing values are (MAR) missing at random.

Denoising Autoencoders we make modifications to the reconstruction cost term so our model does not begin to overfit. In DAE we make a copy of our desired input and instead of fitting that through our model directly we introduce some noise into our input. Thus to make a well generalizable model it not only has to make learn the mapping from input to output but also to be able to remove all the noise which we introduced. For the purpose of this project, we have used the MIDA implementation of denoising autoencoders [45]. **Generative Adversarial Network** developed by Ian Goodfellow has been used for a considerable amount of time to generate new images from previous data. In the last few years, many variations for the same

have been developed which can help us in filling missing values. The outline of GAN's is that there are two neural networks - generator and discriminator [49]. The generator learns on the data and creates new images from the data which gets sent to the discriminator, based on the prediction done by the discriminator the model weights get updated. As we said there exist many implementations for Missing values in GAN we have used GAIN implementation[37]. In **Variational AutoEncoder** we can learn the smooth latent space representations of our original data, this ensures that at every point in the latent space there is a probability associate which relates to something in data. This guarantees that we can every point in the latent space does represent some of our initial data.

If we just try to minimize the reconstruction error by trying the KL divergence term between the prior and the latent space to have similar distribution we would just be copying the initial data and we would have failed to understand the underlying distribution. Variational Autoencoders try to act in the middle of these two extremes by trying to drive the KL divergence to be as small as possible while trying to generate a smooth latent distribution for the data. When we try to optimize both of these constraints we are essentially keeping KL divergence as a regularisation force.

8. Results and Conclusions

The objective of performing this model was to understand the best way we can predict the values for INR. In most research only coagulation parameters are used for this purpose there is a reason for this, there is a lot of research which have proved comprehensively the role coagulation parameters to predict INR. The problems with performing data-driven predictions directly using just coagulation factors are the extent of outliers and the number of missing values in the data. Even though the MIMIC 3, which is one of the most comprehensive medical datasets currently available, even it has coagulation factor values for less than 1% of the number of patients. Thus, there is an increasing risk that, even if there were a very small fraction of points as outliers because we are using this data to generate the values for the missing data, we might actually be making the problem even worse. To mitigate this problem I have proposed using the information, we will get from other parts of datasets, i.e the biochemical parameters. To confirm the results, we have used both the domain knowledge to weed out incorrect corrections[50].

The output label for this problem was International Normalised Ratio (INR) for the database in all three stages of the problem we had more than 95% of the INR values. Since we already have the true value of the INR and assuming that these values are not incorrect the imputed should come as close as possible to these values. We don't expect them o be the same but I expect correct imputations should have mean INR at least in the same order. To check which method comes close to these values, I calculated the mean of the INR data. In stage 1 we decided to reject the database structure even though the overall trends which we obtained were consistent with the literature because the INR was 54. The mean INR of our un-imputed data was 1.34 also another red flag in this imputation had been the R^2 value which showed an abnormal value of 0.99. Unnaturally high INR value in a complex system gave an indication that our model imputation was way off which we confirmed by the INR values. Similar is the case with the mean INR values of stage 2 imputation showed a value of 37. which was again very off from where we expect it to be. In stage 3 we decided to accept these results because the mean INR results we got are 1.5 very close to the actual mean INR values.

9. Future Work

Even though MIMIC data is a time-series database in this work I have not been able to fully take advantage of the sequential data. This is because some of the data did not have a time label associated with it, so I decided to drop it in stage 1 and stage 3 altogether and used a pseudo time series arrangement for stage 2. I relied on features weights from linear regression in order to confirm some aspects of model feasibility, in future, I would like to use a different metric which would not make such strong assumptions about data distribution. Even though INR is accurate in some cases but it is high because of its dependence on a wide number of variables which makes it susceptible to different interpretations by people from diverse experiences. Thus there is a need to make a robust indicator of our body's coagulation performance.

References

[1] S. Palta, R. Saroa, A. Palta, Overview of the coagulation system, Indian journal of anaesthesia 58 (5) (2014) 515.

[2] J. Mosterd, H. Thijssen, The relationship between the vitamin k cycle inhibition and the plasma anticoagulant response at steady-state s-warfarin conditions in the rat., Journal of Pharmacology and Experimental Therapeutics 260 (3) (1992) 1081–1085.

[3] T. Khan, H. Wynne, P. Wood, A. Torrance, C. Hankey, P. Avery, P. Kesteven, F. Kamali, Dietary vitamin k influences intra-individual variability in anticoagulant response to warfarin, British journal of haematology 124 (3) (2004) 348–354.

[4] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3 (2016) 160035.

[5] I. Pratama, A. E. Permanasari, I. Ardiyanto, R. Indrayani, A review of missing values handling methods on time-series data (2016) 1–6.

[6] J. Horton, B. Bushwick, Warfarin therapy: evolving strategies in anticoagulation., American family physician 59 (3) (1999) 635–646.

[7] S. Schulman, S. Granqvist, M. Holmström, A. Carlsson, P. Lindmarker, P. Nicol, S.-G. Eklund, S. Nordlander, G. Lärfars, B. Leijd, et al., The duration of oral anticoagulant therapy after a second episode of venous thromboembolism, New England Journal of Medicine 336 (6) (1997) 393–398.

[8] P. W. Majerus, Anticoagulant, thrombolytic, and antiplatelet drugs, Goodman and Gilman's the pharmacological basis of therapeutics (2001) 1519–1538.

[9] P. Fan, Y. Gao, M. Zheng, T. Xu, P. Schoenhagen, Z. Jin, Recent progress and market analysis of anticoagulant drugs, Journal of thoracic disease 10 (3) (2018) 2011.

[10] M. Kuruvilla, C. Gurk-Turner, A review of warfarin dosing and monitoring, Proceedings (Baylor University. Medical Center) 14 (3) (2001) 305.

[11] A. López Jaimes, C. A. Coello Coello, D. Chakraborty, Objective reduction using a feature selection technique, in: Proceedings of the 10th annual conference on Genetic and evolutionary computation, ACM, 2008, pp. 673–680.

[12] M. K. Sohrabi, A. Tajik, Multi-objective feature selection for warfarin dose prediction, Computational Biology and chemistry 69 (2017) 126–133.

[13] S. Aidit, Y. C. Soh, C. S. Yap, T. M. Khan, C. F. Neoh, S. Shaharuddin, Y. W. Kassab, R. P. Patel, L. C. Ming, Effect of standardized warfarin treatment protocol on anticoagulant effect: comparison of a warfarin medication therapy adherence clinic with usual medical care, Frontiers in pharmacology 8 (2017) 637.

[14] B. O. Sonuga, D. A. Hellenberg, C. S. Cupido, C. Jaeger, Profile and anticoagulation outcomes of patients on warfarin therapy in an urban hospital in cape town, south africa, African journal of primary health care & family medicine 8 (1) (2016) 1–8.

[15] E. K. Kabagambe, T. M. Beasley, N. A. Limdi, Vitamin k intake, body mass index and warfarin maintenance dose, Cardiology 126 (4) (2013) 214–218.

[16] L. Yang, W. Ge, F. Yu, H. Zhu, Impact of vkorc1 gene polymorphism on interindividual and interethnic warfarin dosage requirement—a systematic review and meta analysis, Thrombosis research 125 (4) (2010) e159–e166.

[17] A. Sharabiani, A. Bress, E. Douzali, H. Darabi, Revisiting warfarin dosing using machine learning techniques, Computational and mathematical methods in medicine 2015.

[18] M. Vecsler, R. Loebstein, S. Almog, D. Kurnik, B. Goldman, H. Halkin, E. Gak, Combined genetic profiles of components and regulators of the vitamin k-dependent γ-carboxylation system affect individual sensitivity to warfarin, Thrombosis and haemostasis 95 (02) (2006) 205–211.

[19] M. J. Rieder, A. P. Reiner, B. F. Gage, D. A. Nickerson, C. S. Eby, H. L. McLeod, D. K. Blough, K. E. Thummel, D. L. Veenstra, A. E. Rettie, Effect of vkorc1 haplotypes on transcriptional regulation and warfarin dose, New England Journal of Medicine 352 (22) (2005) 2285–2293.

[20] E. Marek, J. D. Momper, R. N. Hines, C. M. Takao, J. C. Gill, V. Pravica, A. Gaedigk, G. J. Burckart, K. A. Neville, Prediction of warfarin dose in pediatric patients: an evaluation of the predictive performance of several models, The Journal of Pediatric Pharmacology and Therapeutics 21 (3) (2016) 224–232.

[21] G. Tan, H. Cohen, F. Taylor, J. Gabbay, Audit of start of anticoagulation treatment in inpatients., Journal of clinical pathology 46 (1) (1993) 67–71.

[22] N. Doble, J. Baron, Anticoagulation control with warfarin by junior hospital doctors, Journal of the Royal Society of Medicine 80 (10) (1987) 627–627.

[23] R. Dyar, S. Hall, B. McIntyre, Warfarin prescription and administration: reducing the delay, improving the safety, BMJ Open Quality 4 (1) (2015) u204509–w1983.

[24] B. Weiner, P. A. Faraci, R. Fayad, L. Swanson, Warfarin dosage following prosthetic valve replacement: effect of smoking history, Drug intelligence & clinical pharmacy 18 (11) (1984) 904–906.

[25] S. Nathisuwan, P. Dilokthornsakul, N. Chaiyakunapruk, T. Morarai, T. Yodting, N. Piriyachananusorn, Assessing evidence of interaction between smoking and warfarin: a systematic review and meta-analysis, Chest 139 (5) (2011) 1130–1139.

[26] K. Bachmann, R. Shapiro, R. Fulton, F. T. Carroll, T. J. Sullivan, Smoking and warfarin disposition, Clinical Pharmacology & Therapeutics 25 (3) (1979) 309–315.

[27] G. Khoury, M. Sheikh-Taha, Effect of age and sex on warfarin dosing, Clinical pharmacology: advances and applications 6 (2014) 103.

[28] R. Renzi, S. Finkbeiner, Ciprofloxacin interaction with sodium warfarin: a potentially dangerous side effect, The American journal of emergency medicine 9 (6) (1991) 551–552.

[29] T.-c. Fu, A review on time series data mining, Engineering Applications of Artificial Intelligence 24 (1) (2011) 164–181.

[30] G. M. Weiss, H. Hirsh, Learning to predict rare events in categorical time-series data, in: Proceedings of the 1998 AAAI/ICML Workshop on Time-Series Analysis, Madison, Wisconsin, 1998.

[31] P. Dirac, The lorentz transformation and absolute time, Physica 19 (1-–12) (1953) 888–896. doi:10.1016/S0031-8914(53)80099-6.

[32] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, Vol. 1, Springer series in

statistics New York, 2001.

[33] D. B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.

[34] J. Salojärvi, K. Puolamäki, S. Kaski, On discriminative joint density modeling, in: J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, L. Torgo (Eds.), Machine Learning: ECML 2005, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 341–352.

[35] K. Bhaskaran, L. Smeeth, What is the difference between missing completely at random and missing at random?, International journal of epidemiology 43 (4) (2014) 1336–1339.

[36] T. W. Taris, A primer in longitudinal data analysis, Sage, 2000.

[37] J. Yoon, J. Jordon, M. Van Der Schaar, Gain: Missing data imputation using generative adversarial nets, arXiv preprint arXiv:1806.02920.

[38] D. A. Bennett, How can i deal with missing data in my study?, Australian and New Zealand journal of public health 25 (5) (2001) 464–469.

[39] W. M. Campion, Book review: Multiple imputation for nonresponse in surveys (1989).

[40] G. Endler, P. Baumgärtel, A. M. Wahl, R. Lenz, Force: Is estimation of data completeness through time series forecasts feasible? (2015) 261–274.

[41] J. Honaker, G. King, What to do about missing values in time-series cross-section data, American Journal of Political Science 54 (2) (2010) 561–581.

[42] A. Ng, Cs229 lecture notes, CS229 Lecture notes 1 (1) (2000) 1–3.

[43] J. Scheffer, Dealing with missing data.

[44] A. Plaia, A. Bondi, Single imputation method of missing values in environmental pollution data sets, Atmospheric Environment 40 (38) (2006) 7316–7330.

[45] L. Gondara, K. Wang, Mida: Multiple imputation using denoising autoencoders, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 260–272.

[46] E. M. Beale, R. J. Little, Missing values in multivariate analysis, Journal of the Royal Statistical Society: Series B (Methodological) 37 (1) (1975) 129–145.

[47] S. Dray, J. Josse, Principal component analysis with missing values: a comparative survey of methods, Plant Ecology 216 (5) (2015) 657–667.

[48] P. Royston, I. R. White, et al., Multiple imputation by chained equations (mice): implementation in stata, J Stat Softw 45 (4) (2011) 1–20.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.

[50] P. Madley-Dowd, R. Hughes, K. Tilling, J. Heron, The proportion of missing data should not be used to guide decisions on multiple imputation, Journal of clinical epidemiology 110 (2019) 63–73.

table table

Table Name	Table Dimensions	Information Type
Lab Events	(27854055,9)	Patient sequential measurement of fluids
Chart Events	(7642110,3)	Patient entering-leaving history, demographic
Output Events	(434921714,14)	Time when medicine tests were administered
Patients	(46520,9)	Information provided during entering leaving time
Input mv	(3618990,31)	Procedures Performed

Table 1: Tables Used in Data Merging Process

Stage No.	Merging Ty pe	Database Shape
Stage 1	Inner Merging on Coagulation and and Biochemical Data	(475,03,13)
Stage 2	Outer Merging on Coagulation and and Biochemical Data	(5 Million,34)
Stage 3	Inner Merging on some Coagulation and Biochemical Data	(61,453,17)

Table 2: Database Shape at different Stages

Factors (x_j)	Coefficient (A_j)
Factor2	-329.5068
Factor9	-49.5517
Factor5	0.7326
Factor7	232.3636
Factor8	0.0883
Factor10	-3.3436
Factor11	-1.3635
Factor12	42.0101
Antithrombin	-0.6836
Thrombin	-2.4305
Large Platelets	-9.5587
Platelets Smear	330.0326

Table 3: Stage 1 feature Importance

Factors (x_j)	**Coefficient** (A_j)
Factor2	6.5194
Factor9	-120.9961
Factor5	0.3936
Factor7	-1.2677
Factor8	2.1431
Factor10	-1.2526
Factor11	-5.3188
Factor12	-12.7276
Antithrombin	0.6362
Thrombin	-0.0494
Large Platelets	138.3660
Platelets Smear	0.00000
Bilirubin	0.0.692
PO_2	-0.-0.0052
Oxygen	0.0041
Urea Nitrogen	-0.0212
WBC	0.0152
WBC Count	0.0279
Bicarbonate	-0.0334
Sodium	-0.0259
Potassium	0.0303
Potassium - Whole Blood	0.0474

Table 4: Stage 2 feature Importance

Factors (x_j)	**Coefficient** (A_j)
Bilrubin	0.0514
Large Platelets	000
Platelets Count	0.000
PO_2	-0.-0.0052
Oxygen	-0.00003817
Urea Nitrogen	0.0045
WBC	0.020
Bicarbonate	-0.0013
Sodium	-0.0028
Potassium	-0.0718
Potassium - Whole Blood	0.1548
Weight	0.0001
Length of Stay	-0.0017
Female Gender	1.4554
Male Gender	1.4261

Table 5: Stage 3 feature Importance

Method	**Root Mean Square Error**
Mean, Median, Mode	1.2403, 1.2395, NA
Multiple Imputation Chained Equations	0.5659
Expectation Maximization	0.6454
MissForest	0.4992
Generative Adverserial Networks	0.007
Denoising Autoencoders	0.004
Variational Autoencoders	0.002

Table 6: Results for Stage 1 Merging

Method	Root Mean Square Error
Mean, Median, Mode	0.8089, 0.8189, NA
Multiple Imputation Chained Equations	0.5343
Expectation Maximization	Did not converge
Generative Adverserial Networks	0.012224
Denoising Autoencoders	0.00034

Table 7: Results for Stage 2 Merging

Method	Root Mean Square Error
Multiple Imputation Chained Equations	0.5659
Expectation Maximization	0.6454
MissForest	0.5837
Generative Adverserial Networks	0.09139
Denoising Autoencoders	0.0009278

Table 8: Results for Stage 3 merging

Merging	Mean INR	Best Performing Method	Minimum RMSE	R^2
Stage 1	54	Variational Autoencoders	0.002	0.99
Stage 2	37	Denoising Autoencoders	0.00034	0.65
Stage 3	1.5	Denoising Autoencoders	0.0009278	0.15

Table 9: Feasibility Study comparing Different Mergings