# PREDICTING SURVIVAL OF PEOPLE WITH HEART FAILURE USING OVERSAMPLING, FEATURE SELECTIONS AND DIMENTIONALITY REDUCTION

Gunukula Niharika
*ginikulaniharika.191it118@nitk.edu.in*
*Department of Information Technology*
*National Institute of Technology Karnataka*
Surathkal, Mangalore

Annam Indhu Lekha
*annamindhulekha.191it207@nitk.edu.in*
*Department of Information Technology*
*National Institute of Technology Karnataka*
Surathkal, Mangalore

Leela Akshaya
*leelaakshaya.191it226@nitk.edu.in*
*Department of Information Technology*
*National Institute of Technology Karnataka*
Surathkal, Mangalore

Anand Kumar M
*m_anandkumar@nitk.edu.in*
*Department of Information Technology*
*National Institute of Technology Karnataka*
Surathkal, Mangalore

***Abstract*—Cardiovascular diseases are deadly and kill millions of people around the world every year. Heart failure is one of the unfortunate consequence where the heart is unable to pump enough blood for the body. A medical checkup of these patients with attributes including creatinine phosphokinase, ejection fraction, serum creatinine and serum sodium can be used for analysis. In this paper, we have analysed this clinical data and built machine learning models that can predict the survival rate of heart failure of a person. We have used various dimensionality reduction techniques to analyse the data with the aim of reducing the dimensions of the dataset. Finally, we reduced the overfitting of data using Synthetic Minority Oversampling Technique(SMOTE) and Adaptive Synthetic(ADASYN).**

***Index Terms*—Dimentionality reduction, Principal Component Analysis, Isometric Mapping, Sample Minority Oversampling Technique, Independent Component Analysis, T- Stochastic Neighbourhood Embedding, Uniform Manifold Approximation and Projection, Synthetic Minority Oversampling Technique, Adaptive Synthetic.**

## I. Introduction

Nowadays with the accelerated pace in life and inactivity most of the people ignore their health. People are working hard to satisfy their needs but they are not able to spend much time for themselves. This is leading to mental and physical stress. There are several reasons for heart failure like high BP, obesity [13], etc. Heart failure is becoming a very serious problem. It became very common in urban areas due to increased physical and mental stress. There are few cases that reports heart failure due to corona virus. Heart failure became one of the most important factor of death in both men and women. If people don't pay attention to this heart failure, it could finally lead to death.

Predicting the heart failure in the early stage saves the life of many people. Population wide strategies can help in the prevention of heart diseases by addressing behavioural risk factors such as the use of tobacco, obesity, unhealthy diet and physical inactivity. There are many existing heart failure prediction methods but those methods have their own limitations. In this paper, we tried to overcome those limitations and predict the heart failure more accurately and rank the features according to the most important risk factors. We have used many Machine Learning and Deep Learning models like SVM [3], Naive bayes [4], Random Forest [5], Decision trees [6], Logistic Regression [7], gradient Boosting [8], ANN. We have used UCI clinical dataset that has 13 features. Serum creatinine and ejection fraction are one of the most important factors to forecast the proportion of survivors. When a muscle breaks down, creatine generates a waste product called serum creatinine. Serum creatinine level is based on a blood test that measures the amount of creatinine in your blood. It tells how good your kidneys are working. The ejection fraction is the percentage of how much blood the left ventricle pumps out with each contraction.

## II. Literature Survey

Davide Chicco & Giuseppe Jurman [10] proposed an Machine learning model which can predict survival of patients with heart failure from medical features like serum creatinine and ejection fraction. They selected these 2 factors namely, serem creatinine and ejection fraction to predict the survival of patients with heart failure. They did not perform any data preprocessing or any oversampling to address the fact that the dataset used is small.

Jing Wang [11] had conducted a comparative study on Heart Failure Prediction with Machine Learning. He compared the performance of 18 different machine learning models for heart failure prediction based on 12 clinical features. He also oversampled the dataset using SMOTE and performed z score normalisation and min max normalisation. But he did not try reducing the number of features in the dataset.

Asif Newaz, Nadim Ahmed, Farhan Shahriyar Haq [12] predicted heart failure using Recursive Feature Selection. Recursive Feature Selection (RFE) is used to select a suitable subset of features from the original list of features that optimizes the survival rate prediction performance. They also

performed Chi-Squared test to identify the features that are highly related to the target variable.

P Pushpavathi, Santhosh Kumari, N K Kubra [19] performed heart failure prediction using Feature Ranking Analysis in Machine Learning. They used various correlation techniques and to see how one feature is related to another feature. Finally they got to their best accuracy using a ML model with Random Forests.

## III. Dataset And Analysis

For our experiment, we analyzed the clinical records of 299 heart failure patients. The dataset was collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) from April–to December 2015. [1] The dataset has information regarding 194 males and 105 females. The age of all the data subjects was in the range of 40 and 95 years. All patients had left ventricular systolic dysfunction and they had previous heart failures that put them in classes III or IV of the New York Heart Association (NYHA) classification of the stages of heart failure. [2]

The dataset used had no null values or missing values. Hence there was not much pre-processing required to done.

We have performed exploratory data analysis in this dataset to understand the underlying relationship between different features of the dataset.

### A. Exploratory Data Analysis

For analysing the categorical variables of our dataset - 'anaemia', 'high blood pressure', 'diabetes', 'sex' and 'smoking', we have plotted bar graphs. We found that 43.14% of all the patients were anemic, 35.12% of them were having high blood pressure, and 32.11% of them were regular smokers, 41.81 of them are having diabetes Fig.4.

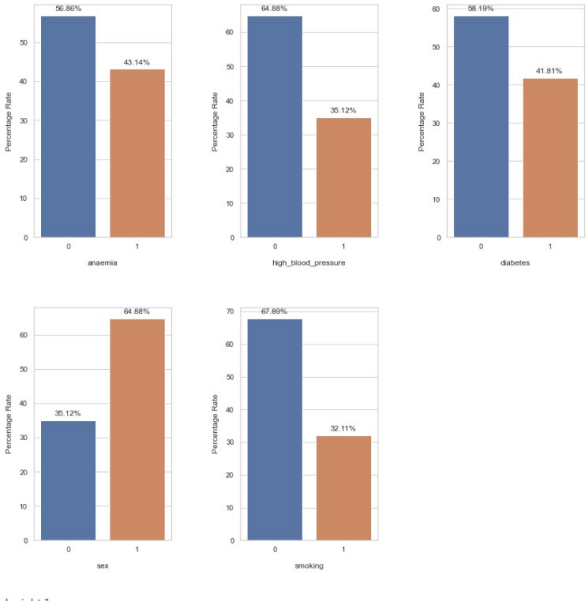


Fig. 1. Analysis of categorical variables

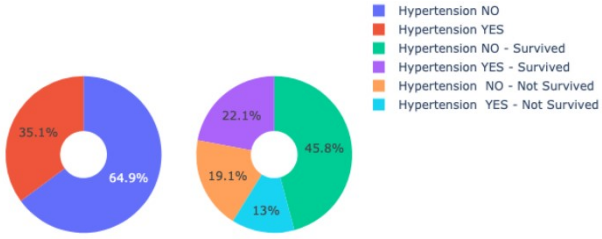


Fig. 2. Effect of High Blood Pressure on survival rate

*1) Effect of High Blood Pressure on survival Rate:* Due to high blood pressure, blood vessels get blocked. This increases the risk of heart failure. So people with high blood pressure are more susceptible to heart diseases. From fig.2, we can say 35% of population are suffering from high blood pressure. Out of these 35%, 22.1% survived the heart failure and remaining 12.9% didn't survive.

In the remaining 65% of population who were not suffering from high blood pressure, 45.8% population survived and remaining 19% didn't survive the heart failure fig.1.

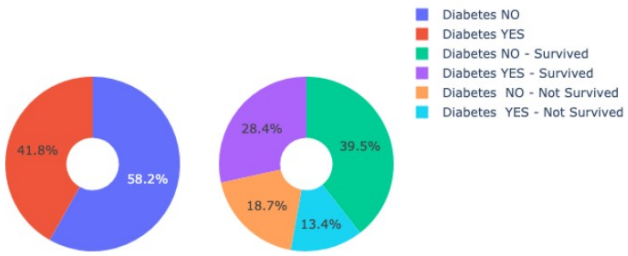


Fig. 3. Effect of Diabetes on survival rate

*2) Effect of Diabetes on survival Rate:* The abnormal handling of glucose and fatty acids by the heart in diabetics increases the risk of heart failure.

The data insights obtained are 41.2% of total population have diabetes and 58.2% don't have diabetes.

Out of these 41.2% of the people who have diabetes, 28.4% of them survived while remaining 13.4% succumb to heart failure. And out of the remaining 58.2% population who don't have diabetes, 39.5% survived and remaining 18.7% succumb to heart failure fig.2.

*3) Effect of Smoking on survival Rate:* Other major risk factor of heart failure is smoking. Because in this condition, plaque builds up in the arteries. This narrows and decreases the blood flow to the heart and also forms blood clots. These blood clots can stop the blood flow partially or completely.

32.1% of the entire population have the habit of smoking. Out of these 32.1%, 22.1% people survived while 10% did not survive. and out of the remaining 67.9% population, 45.8% survived and 22.1% succumb to heart failure fig.3.

TABLE I
SUMMARY OF DATASET [22]

| Feature | Description | Type |
|---|---|---|
| **Age** | Age of the patient | Integer |
| **Anaemia** | Decrease of red blood cells or hemoglobin | Boolean |
| **High blood pressure** | If the patient is suffering from hypertension | Boolean |
| **Creatinine phosphokinase** | Level of the CPK enzyme in the blood | Real |
| **Diabetes** | If the patient has diabetes | Boolean |
| **Ejection Fraction** | Percentage of blood leaving the heart at each contraction | Real |
| **Platelets** | Platelets in the blood | Real |
| **Sex** | Male or Female | Binary |
| **Serum Creatinine** | Level of serum creatinine in the blood | Real |
| **Serum Sodium** | Level of serum sodium in the blood | Real |
| **Smoking** | If the patient smokes or not | Boolean |
| **Time** | Follow-up period - days | Integer |
| **Death Event** | If the patient deceased during the follow-up period | Boolean |

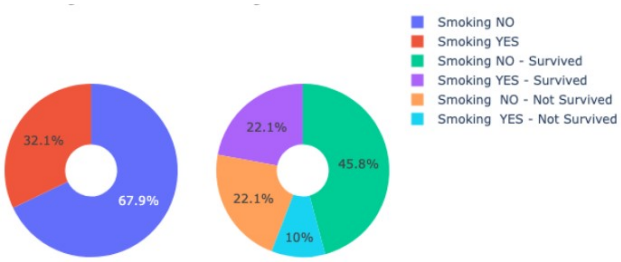


Fig. 4. Effect of Smoking on survival rate

*4) Effect of Creatinine Phosphokinase on survival Rate:* Creatinine Phosphokinase is an enzyme mainly present in heart, brain and skeletal muscles. If the Creatinine phosphokinase levels are very high, then it means that there is an injury or stress to the heart or brain.

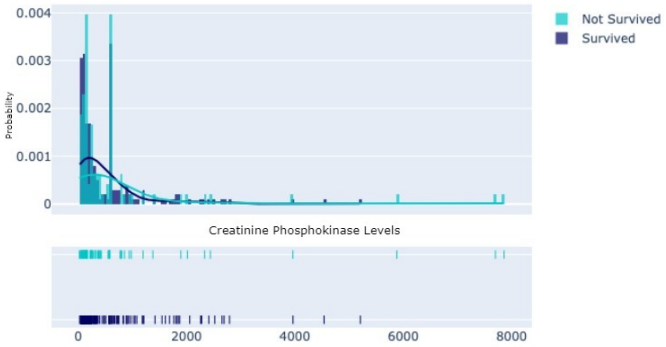


Fig. 5. Effect of Creatinine Phosphokinase on survival rate

From fig.5, we can see that the Creatinine phosphokinase levels are very high in the patients who did not survive the heart failure compared to the patients who survived. In some cases the levels are abnormally high.

*5) Effect of Ejection Fraction on survival Rate:* Ejection fraction tells how well your left or right ventricle pumps blood with each heart beat. Normal Ejection Fraction is about 50 - 75 %

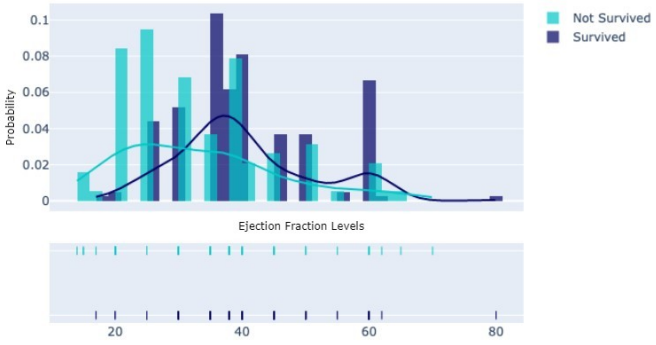


Fig. 6. Effect of Ejection Fraction on survival rate

From fig.6, we can see that most of the people who succumbed to heart failure had less than normal Ejection Fraction i.e., 25-40%

*6) Effect of Serum Creatinine on survival Rate:* Serum Creatinine is the waste product of the blood that comes from the muscles. The normal range of Serum Creatinine for men is 0.75 - 1.35 mg/dL and for woman is 0.59 - 1.04 mg/dL.

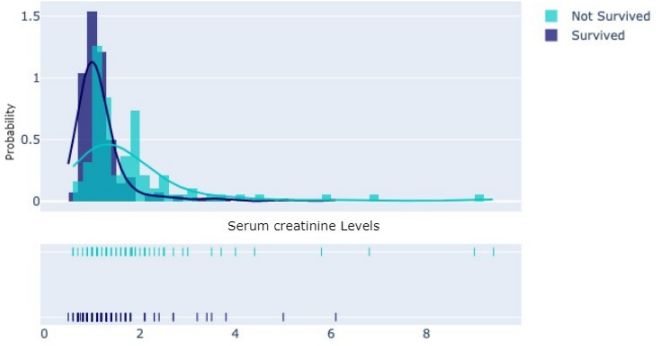


Fig. 7. Effect of Serum Creatinine on survival rate

72.9 % population have very high levels of blood serum and remaining 27.1% have normal serum creatinine levels.

out of these 72.9% population, 43.8% people survived the heart failure and remaining 29.1% succumb to heart failure fig.7.

Out of 27.1% population who have normal range of serum creatinine, a very small number of 3.01% cases did not survive the heart failure condition.

*7) Effect of Serum Sodium on Survival Rate:* The condition of very low Serum Sodium levels is called Hypernatremia. It is a electrolyte disorder and it is mostly seen in people with advanced heart failure. The normal range of Serum Sodium is 135-147 mmol/L.

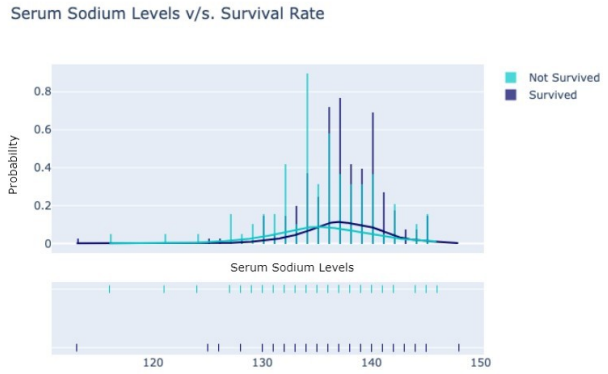


Fig. 8. Effect of Serum Sodium on survival rate

46.5% cases of the total population studied have a very low serum sodium levels and remaining 53.5% cases have acceptable serum sodium levels. Out of these 46.5% cases, 26.8% survived the heart failure and remaining 19.7% did not survive.

Out of the cases having normal range of serum sodium levels 53.5%, 41.1% survived while the remaining 12.4% succumbed to heart failure fig.8.

*8) Follow up Period:* It is the Monitoring period of a person's health after treatment. Majority of the person who

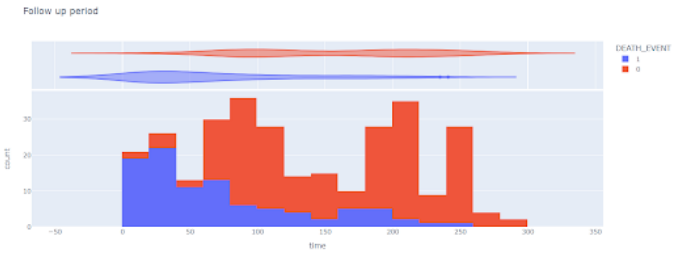


Fig. 9. Effect of Follow up period on survival rate

succumbed to the heart failure condition had less follow up period compared to persons who survived fig.9.

After the Exploratory Data Analysis we have found that the following five features are most important in predicting the heart failure.

1) Age
2) Ejection Fraction
3) Serum Creatinine
4) Serum Sodium
5) Follow up period(Time)

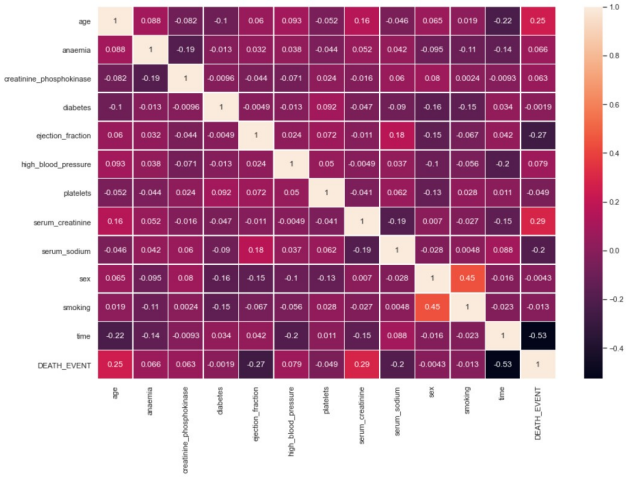


Fig. 10. correlation between all the features and DEATH EVENT

ejection_fraction, age and serum_creatinine features have covariance $>$ 0.2 with feature 'DEATH_EVENT'. Hence we have used all these features only for further analysis.

## IV. Models

1) Support vector machine [3]- It is a non probabilistic linear classifier. This supervised learning algorithm usually gives very accurate results even for high dimensional data. We have used different kernels
   - Linear kernel
   - Radial basis function kernel
   - Sigmoid kernel

   Kernel Function transforms the dataset so that a non-linear decision surface can transform a linear equation to high dimensions.
2) Decision Tree [6] - It is a non parametric supervised learning classifier that depends on structural mapping of binary decisions. Decisions are taken in such a way that the information gain is maximum after the split. We used 2 criteria for this purpose:
   - Entropy
   - Gini

   We have also used **Gradient Boosting** [8] which is basically an ensemble model of many weak decision trees.
3) Logistic Regression [7] - It is a simple statistical method to predict binary outcome
4) Naive Bayes [4] - It is a classifies which is based on Bayes Theorem. It assumes that all features are independent of each other.

TABLE II
RESULTS USING 13 FEATURES FROM ORIGINAL DATASET

| Model Used | Accuracy | Precision | Recall | F1 - Score | ROC_AUC Score |
|---|---|---|---|---|---|
| **SVM - RBF kernal** | 0.68 | 0 | 0 | 0 | 0.5 |
| **SVM - sigmoid kernal** | 0.64 | 0.41 | 0.24 | 0.31 | 0.54 |
| **Decision Tree - entropy criterion** | 0.81 | 0.75 | 0.62 | 0.68 | 0.76 |
| **Decision Tree - gini criterion** | 0.79 | 0.71 | 0.59 | 0.64 | 0.74 |
| **Random Forest - entropy criterion** | 0.87 | 0.87 | 0.67 | 0.78 | 0.82 |
| **Random Forest - gini criterion** | 0.88 | 0.87 | 0.72 | 0.80 | 0.84 |
| **gradient boosting** | 0.89 | 0.88 | 0.76 | 0.81 | 0.85 |
| **Logistic regression** | 0.84 | 0.86 | 0.62 | 0.72 | 0.79 |

TABLE III
RESULTS USING 5 FEATURES FROM ORIGINAL DATASET

| Model Used | Accuracy | Precision | Recall | F1 - Score | ROC_AUC Score |
|---|---|---|---|---|---|
| **SVM - RBF kernal** | 0.84 | 0.89 | 0.59 | 0.71 | 0.78 |
| **SVM - sigmoid kernal** | 0.77 | 1.00 | 0.28 | 0.43 | 0.64 |
| **Decision Tree - entropy criterion** | 0.82 | 0.78 | 0.62 | 0.69 | 0.77 |
| **Decision Tree - gini criterion** | 0.74 | 0.62 | 0.52 | 0.57 | 0.68 |
| **Random Forest - entropy criterion** | 0.88 | 0.85 | 0.76 | 0.8 | 0.85 |
| **Random Forest - gini criterion** | 0.89 | 0.85 | 0.8 | 0.82 | 0.86 |
| **Gradient boosting** | 0.86 | 0.79 | 0.79 | 0.77 | 0.83 |
| **Logistic regression** | 0.84 | 0.8 | 0.69 | 0.74 | 0.80 |
| **Naive bayes** | 0.78 | 0.74 | 0.48 | 0.58 | 0.70 |

TABLE IV
RESULTS USING 13 FEATURES AFTER OVER SAMPLING OUR DATASET USING SMOTE

| Model Used | Accuracy | Precision | Recall | F1 - Score | ROC_AUC Score |
|---|---|---|---|---|---|
| **SVM - RBF kernal** | 0.71 | 0.67 | 0.21 | 0.32 | 0.58 |
| **SVM - Sigmoid kernal** | 0.46 | 0.29 | 0.48 | 0.36 | 0.46 |
| **Decision tree - entropy criterion** | 0.78 | 0.70 | 0.55 | 0.62 | 0.72 |
| **Decision tree - gini criterion** | 0.84 | 0.78 | 0.72 | 0.75 | 0.81 |
| **Random forest - entropy criterion** | 0.86 | 0.77 | 0.80 | 0.78 | 0.84 |
| **Random forest - gini criterion** | 0.89 | 0.85 | 0.79 | 0.82 | 0.86 |
| **Logistic regression** | 0.77 | 0.61 | 0.79 | 0.68 | 0.77 |
| **Naive bayes** | 0.81 | 0.69 | 0.76 | 0.72 | 0.79 |

### *A. Dimensionality Reduction techniques*

As it is harder to work and visualize data with high dimensions, we use dimensionality reduction to remove redundant and correlated data and reduce the dimensions of the dataset. We have applied the following 5 dimensionality reduction techniques with the aim of better visualization.

1) Principal Component Analysis [14]
2) Fast ICA [15]
3) Isomap [16]
4) T-SNE [17]
5) UMAP [18]

*1) Principal Component Analysis:* PCA is a linear dimensionality reduction technique. The data in higher dimension is projected on to the lower dimensional space in such a way that it maximizes the variance in lower dimensions.

we have reduced the dimensions of our dataset to five features using PCA, trained using different models and analyzed the accuracy of these models.

### *B. Oversampling the dataset*

*1) SMOTE:* Synthetic Minority Oversampling Technique [18] is a oversampling technique where it uses k - nearest neighbour algorithm to generate synthetic data.

*2) ADASYN:* Adaptive Synthetic Sampling [19] approach is a oversampling technique that uses weighted distribution for different minority class samples according to their difficulty level in learning and generates more synthetic data. we have used these oversampling techniques on our dataset

## V. RESULTS AND ANALYSIS

Out of all the metrics, we are keen on improving the Recall. This is because we would not want to miss out on any positive cases of death due to heart failure.

We have calculated the following metrics for every model:

- Accuracy - It is the percentage of total correct predictions
- Precision - It says what percentage of all the positives predicted are truly positive.
- Recall - It represents the true positive rate. I.e. out of all the positives in the data, what percentage of them are predicted successfully.

TABLE V
RESULTS USING 5 FEATURES AFTER OVER SAMPLING OUR DATASET USING SMOTE

| Model Used | Accuracy | Precision | Recall | F1 - Score | ROC_AUC Score |
|---|---|---|---|---|---|
| **SVM - RBF kernal** | 0.89 | 0.85 | 0.79 | 0.82 | 0.86 |
| **SVM - Sigmoid kernal** | 0.63 | 0.44 | 0.48 | 0.46 | 0.60 |
| **Decision tree - entropy criterion** | 0.74 | 0.61 | 0.59 | 0.60 | 0.70 |
| **Decision tree - gini criterion** | 0.84 | 0.8 | 0.69 | 0.74 | 0.80 |
| **Random forest - entropy criterion** | 0.87 | 0.79 | 0.79 | 0.79 | 0.85 |
| **Random forest - gini criterion** | 0.84 | 0.76 | 0.76 | 0.76 | 0.82 |
| **Logistic regression** | 0.8 | 0.65 | 0.83 | 0.73 | 0.81 |
| **Naive bayes** | 0.84 | 0.76 | 0.76 | 0.76 | 0.82 |

TABLE VI
RESULTS AFTER REDUCING DIMENSIONS TO 5 FEATURES USING PCA AND OVER SAMPLING OUR DATASET USING SMOTE

| Model Used | Accuracy | Precision | Recall | F1 - Score | ROC_AUC Score |
|---|---|---|---|---|---|
| **SVM - RBF kernal** | 0.56 | 0.37 | 0.52 | 0.43 | 0.55 |
| **SVM - Sigmoid kernal** | 0.53 | 0.38 | 0.72 | 0.5 | 0.58 |
| **Decision tree - entropy criterion** | 0.84 | 0.76 | 0.76 | 0.76 | 0.82 |
| **Decision tree - gini criterion** | 0.8 | 0.70 | 0.66 | 0.68 | 0.76 |
| **Random forest - entropy criterion** | 0.81 | 0.69 | 0.76 | 0.72 | 0.78 |
| **Random forest - gini criterion** | 0.82 | 0.69 | 0.79 | 0.74 | 0.81 |
| **Logistic regression** | 0.76 | 0.58 | 0.90 | 0.70 | 0.79 |
| **Naive bayes** | 0.84 | 0.76 | 0.76 | 0.76 | 0.82 |

TABLE VII
RESULTS USING 5 FEATURES AND OVER SAMPLING OUR DATASET USING ADAPTIVE SYNTHETIC SAMPLING APPROACH

| Model Used | Accuracy | Precision | Recall | F1 - Score | ROC_AUC Score |
|---|---|---|---|---|---|
| **SVM - RBF kernel** | 0.83 | 0.70 | 0.83 | 0.76 | 0.83 |
| **SVM - sigmoid kernel** | 0.57 | 0.38 | 0.51 | 0.43 | 0.55 |
| **Descision tree - entropy criteria** | 0.86 | 0.79 | 0.76 | 0.77 | 0.83 |
| **Descision tree - gini criteria** | 0.83 | 0.75 | 0.72 | 0.74 | 0.80 |
| **Random forest - entropy criteria** | 0.89 | 0.82 | 0.82 | 0.82 | 0.87 |
| **Random forest - gini criteria** | 0.84 | 0.74 | 0.79 | 0.77 | 0.83 |
| **Gradient boosting** | 0.86 | 0.77 | 0.79 | 0.78 | 0.84 |
| **Logistic regression** | 0.8 | 0.65 | 0.82 | 0.72 | 0.80 |
| **Naive bayes** | 0.82 | 0.72 | 0.72 | 0.72 | 0.80 |

TABLE VIII
RESULTS AFTER USING NEURAL NETWORKS

| Dataset | Accuracy | Precision | Recall | F1 - Score | ROC_AUC Score |
|---|---|---|---|---|---|
| **13 features** | 0.81 | 0.83 | 0.52 | 0.64 | 0.73 |
| **5 features** | 0.88 | 0.78 | 0.86 | 0.82 | 0.87 |
| **SMOTE + 5 features** | 0.84 | 0.74 | 0.79 | 0.77 | 0.83 |
| **SMOTE + 13 features** | 0.36 | 0.33 | 1 | 0.5 | 0.52 |
| **ADASYN + 5 features** | 0.9 | 0.81 | 0.90 | 0.85 | 0.90 |
| **ADASYN + 13 features** | 0.69 | 0.6 | 0.10 | 0.18 | 0.54 |

- F1 - Score - It is the harmonic mean of both precision and recall. It takes false positives and false negatives into account.
- ROC_AUC Score - It is the area under the probability curve that plots TPR (True positive rate) against FPR (False positive rate). In electrical terms, it separates the signal from the noise. The higher the value, the better the model separates positives from negative.

### *A. Predictions using 13 features*

We have used various models as described before on all the 13 features of our dataset. The results are documented in Table III. We achieved good accuracy for decision tree and random forest with entropy criteria for making splits. And as expected even gradient boosting performed very well. But the recall scores that we should primarily focus on are not quiet satisfactory, with the best being 0.75 from gradient boosting.

### *B. Predictions using 5 features*

We have selected 5 features from 13 features concluded from the exploratory data analysis. We have used all models the that we used previously and have documented the results in Table IV. We could not increase the total accuracy but we

TABLE IX
BEST PERFORMANCE METRIC SCORES FOR EVERY MODIFIED DATASET USING MACHINE LEARNING MODELS

| Dataset | Best Model | Accuracy | Precision | Recall | F1 - Score | ROC_AUC Score |
|---|---|---|---|---|---|---|
| **13 features** | Gradient boosting | 0.89 | 0.88 | 0.76 | 0.81 | 0.85 |
| **5 features** | Random Forest - Gini criterion | 0.89 | 0.85 | 0.79 | 0.82 | 0.86 |
| **SMOTE + 5 features** | SVM - RBF kernel | 0.89 | 0.85 | 0.79 | 0.82 | 0.86 |
| **SMOTE + 13 features** | Random Forest - Gini criterion | 0.89 | 0.85 | 0.79 | 0.82 | 0.86 |
| **SMOTE + 5 features using PCA** | Logistic Regression | 0.75 | 0.58 | 0.90 | 0.70 | 0.79 |
| **ADASYN + 5 features** | Random Forest - Entropy criterion | 0.89 | 0.83 | 0.83 | 0.83 | 0.87 |

achieved a recall score of 0.79 which is approximately 0.04 higher than the results obtained from 13 features.

### *C. Predictions using 13 features after over sampling our dataset using SMOTE*

We have performed SMOTE oversampling technique on our original dataset to oversample the minority classes. We have trained various models using this oversampled dataset and have documented the results in Table V. The best accuracy that we obtained in this method is the same as the best accuracy obtained using 5 features, ie 0.88 for Random forest model using Gini criteria for making the split at decision boundries. We also obtained good accuracy of 0.84 and and 0.87 for Decision Tree using Gini criteria and Random Forest using Entropy criteria respectively. We also obtained same recall of 0.79 using Random Forest and Logistic Regression.

### *D. Predictions using 5 features after over sampling our dataset using SMOTE*

We selected 5 features from 13 features concluded from exploratory data analysis. We also oversampled our dataset using SMOTE and used all the models that we have used previously. The results are documented in Table VI.

We did not find any better accuracy compared to previous techniques, but obtained 0.82 recall, which is higher than the best obtained using the 13 feature SMOTE oversampled dataset.

### *E. Predictions using SMOTE oversampled dataset and reduced dimensions to 5 features using PCA*

Using SMOTE oversampling and dimension reduction using PCA we were able to specifically tune the models to give a better recall score, even though the accuracy score has not increased. This is a desired result as recall score is a more important metric for our problem statement. We obtained the best recall score of 0.89 for logistic regression. All the results for this modified dataset is documented in Table VII.

### *F. Predictions using 5 selected features and ADASYN oversampled dataset*

We have selected 5 features from 13 features concluded using exploratory data analysis, and performed Adaptive Synthetic sampling approach for sampling of imbalanced class. We made predictions using all the models used previously and have documented our results in Table VIII. The best accuracy obtained is 0.86 for Random Forest using entropy criteria and the best recall obtained is 0.86 for logistic regression.

### *G. Predictions using Neural Networks*

We used a small neural network, with 11 neurons and a relu activation function. Then using the sigmoid function, we calculated the probability of death due to heart failure.

Table X and Table XI summarize the best performance metrics scores for every modified dataset.

TABLE X
COMPARISON WITH EXISTING WORKS

| Paper-Model | Precision | Recall | Accuracy | F1 - Score |
|---|---|---|---|---|
| Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone [10]: Best model- Descision Tree | 0.532 | 0.491 | 0.740 | 0.547 |
| Heart Failure Prediction with Machine Learning [11]: Best Model - Linear regression with Z score normalisation of dataset | - | - | 0.8667 | 0.8333 |
| Survival prediction of heart failure patients using machine learning techniques [12]: Best ML model - Logistic Regression | - | 0.34 | 0.74 | - |
| Heart Failure Prediction by Feature Ranking Analysis in Machine Learning [19]: Best ML model - Random Forest | - | - | 0.816 | - |
| Our Neural Network - reduced dataset with 5 features | 0.81 | 0.8965 | 0.9 | 0.8524 |
| Our ML Model - Random Forest with oversampled dataset using ADASYN | 0.83 | 0.8275 | 0.88 | 0.8275 |

## VI. CONCLUSION

We have modified the dataset in various ways. We used 5 features selected using exploratory data analysis. We explored dimentionality reduction using PCA. To overcome the problem of have a small dataset, we used and compared 2 oversampling techniques, SMOTE and ADASYN. Using the new modified dataset, we tried different machine learning models like SVM, decision trees, random forests and also a naive bayes classifier.

The best machine learning model that we obtained had 0.88 accuracy, 0.82 recall and 0.82 F-1 score. This performance was

obtained with a reduced dataset of 5 features, oversampled with ADASYN. This is a small but significant improvement from the experiments conducted by other researchers as shown in Table XI.

In the neural network model, using 11 neurons on the same modified dataset performed well with an accuracy of 0.90, recall of 0.89 and a F-1 score of 0.85.

We have experimented with a small neural network, with 1 layer and 11 neurons. We can add more layers and try different architectures and see how the modified dataset behaves. Since deep neural networks need large datasets to train, we can either collect more data that or combine results from different ovesampling techniques.

## REFERENCES


[1] Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. *Survival analysis of heart failure patients: a case study*, PLOS ONE, July 20, 2017.

[2] Bredy C, Ministeri M, Kempny A, Alonso-Gonzalez R, Swan L, Uebing A, Diller G-P, Gatzoulis MA, Dimopoulos K. *New York Heart Association (NYHA) classification in adults with congenital heart disease: relation to objective measures of exercise and outcome*. European Heart Journal - Quality of Care and Clinical Outcomes, 17 August 2017

[3] Boser, Guyon, Vapnik, *Support Vector Machines*, Springer Link, 1992

[4] Vikramkumar, Vijaykumar B, Trilochan: *Bayes and Naive Bayes Classifier* 3 Apr 2014

[5] Leo Breiman , *Random Forests*, Springer Link, October 2001

[6] J. R. Quinlan, *Induction of decision trees*, Springer Link, March 1986

[7] Chao-Ying Joanne, Peng Kuk Lida Lee, Gary M. Ingersoll, *An Introduction to Logistic Regression Analysis and Reporting*, The journal of Educational Research, September 2002

[8] Jerome H. Friedman, *Greedy function approximation: A gradient boosting machine*, Institute of Mathematical Statistics, October 2001

[9] Enzo Grossi, Massimo Buscema,*Introduction to artificial neural networks*, European Journal of Gastroenterology Hepatology, January 2008

[10] Davide Chicco, Giuseppe Jurman, *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone*, BMC Part of Springer, 3rd Feb 2020.

[11] Jing Wang, *Heart Failure Prediction with Machine Learning: A Comparative Study*, Journal of Physics: Conference Series, 2021

[12] Asif Newaz, Nadim Ahmed, Farhan Shahriyar Haq, *Survival prediction of heart failure patients using machine learning techniques*, Elsevir, October 2021

[13] World Health Organization, Cardiovascular Diseases, WHO, Geneva, Switzerland, 2020

[14] Sidharth Mishra, Uttam Sarkar, Subhash Taraphder, Sanjoy Datta, *Principal Component Analysis*, International Journal of Livestock Research, January 2017

[15] Aapo Hyvärinen and Erkki Oja, *Independent Component Analysis: Algorithms and Applications*, John Wiley & Sons, June 2001

[16] Joshua B. Tenenbaum, Vin de Silva, John C. Langford, *A Global Geometric Framework for Nonlinear Dimensionality Reduction*, December 2000

[17] Laurens van der Maaten, Geoffrey Hinton: *Visualizing Data using t-SNE*, Journal of Machine Learning Research, November 2008

[18] Leland McInnes, John Healy, James Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, Arxive Cornell University, Feb 2018

[19] P Pushpavathi, Santhosh Kumari, N K Kubra, *Heart Disease Prediction by Feature ranking Analysis*, International Journal of Advanced Computer Science and Applications, Feb 2021

[20] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. *SMOTE: Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, June 2002

[21] Shutao Li, Edwardo A. Garcia, Yang Bai, Haibo He,IEEE Xplore, *ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning* July 2008

[22] *Heart failure clinical records Data Set*, https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records