**Medical Health Conditions**

**1. Introduction**

Every year, millions of individuals are impacted by the global health catastrophe known as cancer. This is approximately one third of all total deaths worldwide [[1](https://pmc.ncbi.nlm.nih.gov/articles/PMC10809869/#:~:text=Introduction-,Cardiovascular%20diseases%20(CVDs)%20are%20the%20leading%20cause%20of%20mortality%20globally,all%20global%20deaths%20%5B1%5D.)]. Nearly 20 million new cases of cancer and 9.7 million fatalities were reported globally in 2022; by 2050, there would likely be an estimated 35 million new cases annually. Breast cancer and lung cancer continue to be the most often diagnosed types. There will be 972,060 new cases among women. Adolescents (ages 15–19) and children (years 0–14) account for a smaller percentage of all diagnoses, with some malignancies such brain tumours and leukaemia being the most prevalent among them. Consult organisations like the CDC and American Cancer Society for more information on cancer statistics, especially those by age group. Early cancer detection greatly improves the chance of a successful course of therapy; nevertheless, present diagnostic techniques frequently have drawbacks, including exorbitant costs, drawn-out processes, and inconsistent outcomes. Despite their effectiveness, genetic testing, biopsy analysis, and medical imaging require a lot of resources and are prone to delays and human error.

These numbers demonstrate the critical need for improved early detection techniques, especially in areas with limited access to medical care. By analysing medical data, machine learning (ML) approaches enable quicker and more accurate diagnoses, which presents intriguing solutions [[2]](https://pmc.ncbi.nlm.nih.gov/articles/PMC6616540/) [[3]](https://data.mendeley.com/datasets/65gxgy2nmg/2). Due to the availability of more data in the field of healthcare and improvements to machine learning methods, risks of various diseases have become predicted as a promising tool. Machine learning can improve the accuracy of cancer detection by using techniques like feature selection to improve the predictive power of models, outlier identification to clean the data, and SMOTE (Synthetic Minority Over-sampling Technique) to solve class imbalance [[4]](https://www.sciencedirect.com/science/article/pii/S2405844024001956#br0180),These models' robustness and dependability are ensured via hyperparameter tuning and cross-validation, providing hope for better cancer care and treatment outcomes [[5]](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)?gad_source=1&gclid=Cj0KCQiArby5BhCDARIsAIJvjIQ1Fz0cIr0Lq9o4ks26JWIh6Kph9fpXNwKuByiw-HrAiUpfnOx0qY4aAisxEALw_wcB) [[6]](https://link.springer.com/article/10.1007/s11023-024-09682-0). Secondly, tuning of the parameters of the ML model to enhance its accuracy without putting too much effort into building an overly complex model is a fundamental problem in this field.

The main idea is to help people to develop a reliable machine learning model of predicting the risk of a cancer based on their health conditions and habits. To do this we want to take the following steps in order to use data driven techniques/ approaches to solve a few important problems in cancer prediction. Decomposition of large data sets into groups of normal and special cases of outliers techniques will enhance the global prediction capability and reduce the distorted effect of outliers. Again to reduce the computational burden and noise while not reducing the predictive ability, feature selection will also be used in order to select the relevant predictors. Several important phases are under consideration of the methodology for cancer prediction. To begin with, the techniques such as imputation or removal are applied in order to address cases when a value is missing in a dataset. In order to reduce their impact, outliers identify and limit them.

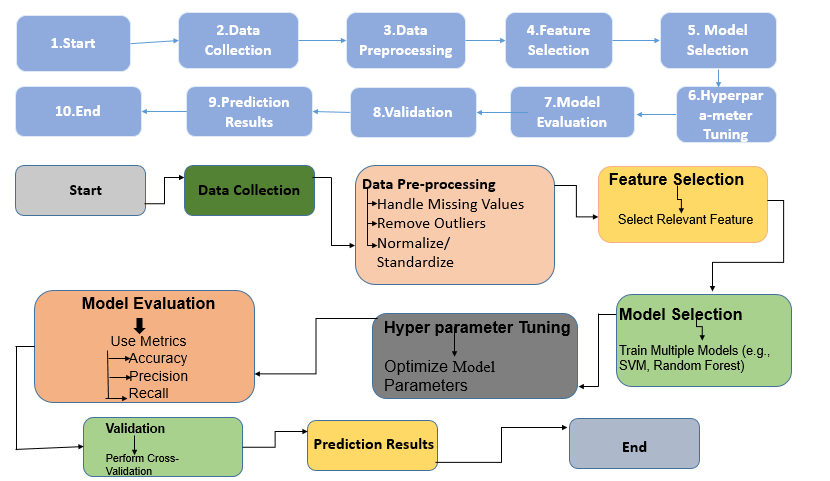
This work provides a very useful input to the field of healthcare improved predictions with features such as balanced and noisy data through the methods of Machine Learning and AI. Risk evaluation gets boosted with methods like SMOTE [[7]](https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/), Outlier Detection [[8]](https://www.sciencedirect.com/topics/engineering/outlier-detection), and the feature selection method [[9]](https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/), which helps healthcare gurus to easily isolate high-risk clients, and apply preventive measures. The study also provides understanding of the magnitude and nature of potential risk factors in identifying individuals and thus contributes to clinical management and population health planning. Lastly, it could help cut down on cancer mortality since the focus can be put on the right individual at the right time, and the right intervention can be delivered.

**2. Related Work**

1. D. Crosby et al. [[10]](https://doi.org/10.1155/2024/5080332) proposed a model for Enhancing cancer Prediction with Machine Learning. They used SVM, RF,DT,Naive Bayes, KNN and they have demonstrated high accuracy rates in predicting cancer disease using machine learning and used feature selection techniques to identify crucial attributes. The Main Gap for this model is not appropriate for a diverse, large dataset and handling missing values. Dataset comprised a total of 486 cases. The dataset is split into 70% for training and 30% for testing. SVM demonstrated an accuracy of 91.8% without PSO, which improved to 94.3% with the incorporation of PSO.
2. N. M. ud din et al. [[11]](https://doi.org/10.1177/14604582241270830) used the Chi-Square Automatic Interaction Detection (CHAID) model to predict mortality in Jordanian men with cancer. They analyzed 3,435 records from Jordan’s electronic health system, focusing on factors like age, pulse oximetry, and systolic blood pressure. Missing biochemical data were excluded, while categorical data were handled directly. The data was split 70:30 for training and testing, with CHAID achieving 93.72% accuracy and highlighting its effectiveness in healthcare prediction.
3. M. Tahmooresi et al. [[12]](https://doi.org/10.3390/electronics13020272) developed an AI-based model using Linear SVM, logistic regression, k-nearest neighbors, and random forest to predict cancer risk in preventive care. Data from 8,763 patient records were preprocessed and split 80:20 for training and testing. Age, heart rate, BMI, and cholesterol emerged as key predictors. Logistic regression demonstrated the best initial screening accuracy for identifying at-risk patients needing further assessment. The model aims to enable preventive strategies and reduce health risks related to modifiable factors.
4. A. E. Pelze et al. [[13]](https://doi.org/10.32604/cmc.2021.014649) proposed a machine learning model combining Support Vector Machine (SVM) and fuzzy-based decision-level fusion to predict cancer disease. The model used two public datasets with 1,025 and 70,000 records, processed with imputation, normalization, and 5-fold cross-validation. Achieving a 96.23% accuracy, the architecture outperformed prior models like hybrid random forest and ANN-based systems. The study shows the potential of SVM-fuzzy fusion in enhancing diagnostic accuracy, providing a viable solution for early heart disease prediction.
5. F. Aigner et al. [[14]](https://doi.org/10.3390/technologies12010004) proposed a Machine Learning-based Congenital cancer Disease Prediction Method (ML-CHDPM) utilizing CNN, BiLSTM, and Attention Mechanisms for ECG signal processing in pregnant women. Data from IoMT sensors on ECG signals was used for classification. Their ML-CHDPM model achieved a 94.28% accuracy, emphasizing advanced signal processing in early CHD detection.
6. A. F. M. Agarap et al. [[15]](https://doi.org/10.3390/a17020078) conducted a comprehensive review on the use of machine learning techniques for cardiovascular disease prediction, covering the benefits, model evaluation, and feature selection methods essential in healthcare. A range of machine learning models, including SVM, random forests, and neural networks, were discussed for their effectiveness. Recent advancements like explainable AI and federated learning were suggested as promising approaches for improving predictive accuracy and patient privacy.
7. T. Tamada et al. [[16]](https://doi.org/10.3390/diagnostics14020144) examined the effectiveness of seven machine learning models (K-Nearest Neighbors, Support Vector Machine, Logistic Regression, CNN, Gradient Boosting, XGBoost, and Random Forest) to predict cancer. Using two datasets with balanced features, they optimized models through data preprocessing, feature scaling, and hyperparameter tuning. XGBoost achieved the best accuracy (98.5%), precision (99.14%), recall (98.29%), and F1 score (98.71%), highlighting its potential in cardiovascular disease diagnosis.
8. Mijwil et al. [[17]](https://doi.org/10.52866/ijcsm.2024.05.01.018) applied seven machine learning algorithms—K-Nearest Neighbors (K-NN), Decision Tree (DT), Linear Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Multilayer Perceptron (MLP), and Random Forest (RF)—to predict cancer disease using a UC Irvine dataset with 303 cases. They split the data into 80% for training and 20% for testing. Among the models, the Multilayer Perceptron (MLP) achieved the highest accuracy, exceeding 88%, while Decision Tree (DT) had the lowest accuracy at 79%, highlighting MLP’s strength in specificity and F1-score.
9. Jalal et al. [[18]](https://doi.org/10.33022/ijcs.v13i2.3921) Proposed cancer Disease Detection Classification Based on Deep Learning Algorithm where he used CNN model along with deep learning model like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) and Generative Adversarial Networks (GANs). They got an accuracy rate of 98%.
10. Verma et al. [[19]](https://doi.org/10.3390/electronics10233013) proposed a secure and efficient disease prediction system integrating IoT and cloud technology. They introduced a novel encryption method, Hybrid Elapid Encryption (HEE), to enhance data security and used a hybrid neuro-fuzzy classifier, GFI-GWALO, for accurate disease prediction and severity assessment. Their system achieved impressive results with 100% accuracy, 99.98% specificity, and a processing time of 8 ms, demonstrating superior security and performance compared to traditional methods in MATLAB simulations.
11. Fortuny et al. [[20]](https://doi.org/10.3390/jcdd9120430) created machine learning models to predict acute myocardial infarction (AMI) before patients reached the hospital, using data from a national EMS database. The study found that machine learning models, particularly extreme gradient boosting (XGB) and multilayer perceptron (MLP), provided better prediction accuracy than traditional methods. The XGB model achieved an AUC of 0.837 and highlighted key predictors like chest pain duration, palpitations, and ST elevation on ECG.
12. Kakadiaris et al. [[21]](https://doi.org/10.1161/jaha.118.009476) created a machine learning-based risk calculator using a Support Vector Machine (SVM) to predict cardiovascular disease (CVD) risk, drawing from data in the MESA cohort. Their model showed high accuracy, with a sensitivity of 86% and specificity of 95%, significantly outperforming the traditional ACC/AHA Risk Calculator, which had a sensitivity of 76% and specificity of 56%. The ML calculator recommended statins for fewer individuals while correctly identifying more high-risk cases, missing fewer CVD events than the conventional approach.
13. Schultz et al. [[22]](https://link.springer.com/chapter/10.1007/978-3-642-13232-2_12#auth-Izabela-Rojek) proposed a Hybrid Neural Networks as Prediction Model to detect cancer. They Used Hybrid Neural Network as a Forecasting Model.
14. Miller et t al. [[23]](https://doi.org/10.3390/app12199596) introduced AI-Based Prediction of Myocardial Infarction Risk as an Element of Preventive Medicine on [Kaggle](https://www.kaggle.com/ronitf/heart-disease-uci/) dataset which contains data on 14 parameters collected from 303 patients studied at 4 European research centers. Used Pandas, Matplotlib, scikit-learn, numpy models initially. Dataset splitted at 80:20 ratio for training and testing. They Used Knn, RF, LR, Linear SVC and got highest accuracy 90% using Logistic Regression. Main gap for this model was their limitations of handling large datasets according to the Device Configurations.

**3. Methodology**

This study paper's methodology uses cutting-edge machine learning approaches to forecast medical conditions for cancer in a methodical and repeatable way. The methods used for data collection, preprocessing, and analysis are described in this part, along with the creation and assessment of predictive models. Throughout the research process, the methodology guarantees ethical compliance and scientific rigour.



**Fig. 1 The proposed workflow for cancer Disease Detection according to our study.**

**3.1 Data Collection**

The study utilises publicly available datasets containing clinical, diagnostic, and demographic information related to cancer. The primary datasets include:**Kaggle Repository**: Cancer datasets featuring patient history, biomarker levels, imaging results, and diagnosis outcomes.**UCI Machine Learning Repository**: Datasets like Breast Cancer Wisconsin (Diagnostic) or similar, which provide comprehensive cancer-related attributes The datasets were selected based on:**Relevance**: Inclusion of significant attributes such as tumor size, histology, molecular markers, patient age, and genetic information.**Class Representation**: Sufficient instances of both positive (cancerous) and negative (non-cancerous) outcomes.**Diversity**

**3.2 Data Preprocessing and Split Dataset**

Preprocessing is a crucial step to improve data quality and ensure compatibility with machine learning algorithms. The following steps were performed:

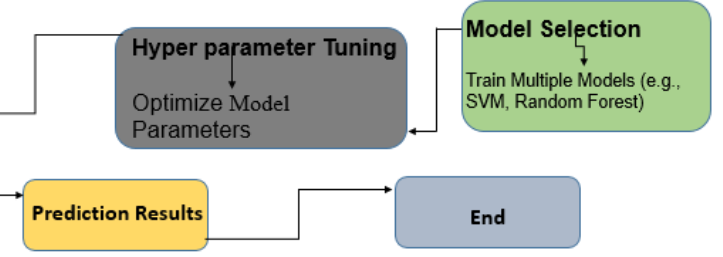
Cleaning: Removed duplicate or irrelevant records. Addressed missing values using imputation techniques such as mean, median, or k-nearest neighbors (KNN).

Outlier Detection and Removal: Utilized statistical methods (e.g., Z-scores, IQR) to detect and remove extreme values that could distort predictions.

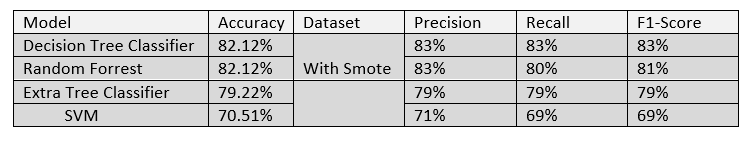
Normalization and Scaling: Applied Min-Max scaling or Standardization to ensure all features are on the same scale, critical for algorithms like SVM or neural networks.

Balancing Class Distribution: Since cancer datasets often have imbalanced classes (e.g., more healthy cases than cancerous cases), Synthetic Minority Oversampling Technique (SMOTE) was used to create a balanced dataset for model training.

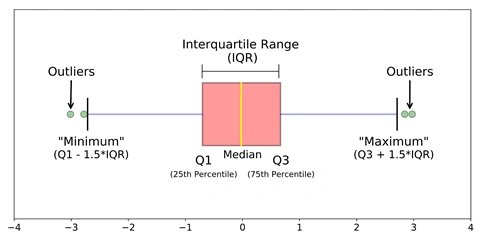
A short Diagram [24]:



**Fig. 2 Valid Dataset Splitting.**

**Table. 1 Dataset without using Smote**

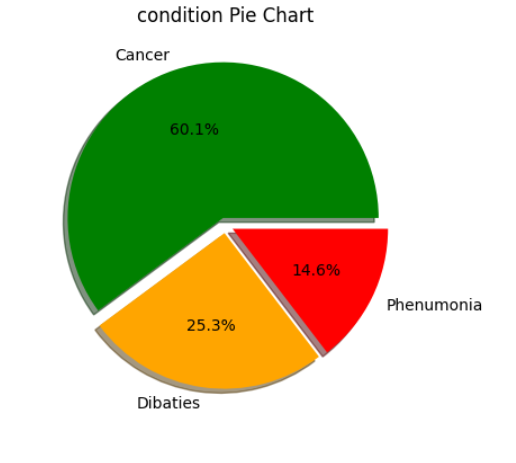
**3.3 Finding Outlier**

Outliers in continuous features like impulse, blood pressure (high and low), glucose, and troponin levels were identified using the interquartile range (IQR) method. Outliers were defined as values below Q1 − 1.5 × IQR or above Q3 + 1.5 × IQR. To reduce their impact while preserving data distribution, values below the lower bound were set to the 1st percentile, and those above the upper bound were capped at the 99th percentile. This approach maintained data integrity and minimized the influence of extreme values on model performance. An overview of outliers in the dataset [26].

**Fig. 3 Outlier Detection Model.**

**3.4 Smote Dataset**

A dataset that has been altered through the use of the Synthetic Minority Oversampling Technique (SMOTE) is referred to as an SMOTE dataset. SMOTE creates synthetic examples for the minority class (the class with fewer instances) in order to balance unbalanced datasets in machine learning. The model may perform poorly on the minority class, therefore this is especially helpful when working with a dataset where one class (for example, cancer-positive patients) contains far fewer examples than the other class (for example, healthy patients). [28].



**Fig. 4 Dataset without Smote.**

**3.4.1 SMOTE (Synthetic Minority Oversampling Technique)**

SMOTE is widely used to handle imbalanced datasets. Instead of duplicating existing minority class examples, it creates new synthetic ones. It does this by choosing a data point from the minority class and then identifying its nearest neighbors. A new point is generated by interpolating between the chosen point and one of its neighbors. This method diversifies the minority class samples, reducing overfitting compared to simply copying the data.

### 

### 

### 

### **3.4.2 Random Oversampler**

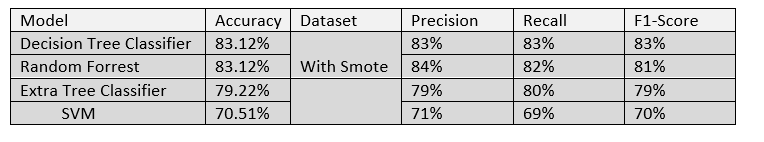
One of the most straightforward methods for dealing with unbalanced datasets in machine learning is the Random Oversampler. By randomly replicating instances from the minority class, Random Oversampling balances the dataset when one class (such as cancer-positive cases) has significantly fewer examples than another (such as healthy cases). This improves the dataset's balance by increasing the amount of minority class samples.

### **3.4.3 ADASYN (Adaptive Synthetic Sampling)**

Adaptive Synthetic Sampling, or ADASYN for short, is a sophisticated machine learning method for dealing with unbalanced datasets. Similar to SMOTE, but with a twist: ADASYN concentrates more on producing synthetic data for cases in the minority class that are more challenging to master. Because of this, it's a clever and flexible way to balance datasets.

### **3.4.4 SMOTE**

Synthetic Minority Oversampling Technique, or SMOTE, is a machine learning technique for dealing with unbalanced datasets. Machine learning models can have trouble with datasets when one class (for example, cancer-positive patients) has a lot less examples than another class (for example, healthy patients). They might start to favour the majority class, which would make it more difficult to predict the minority class accurately.

** Table. 2 Dataset using Smote**

**3.5 Feature selection**

Illness depending on the findings of different medical tests. Numerous features, including blood pressure, cholesterol, age, weight, and even irrelevant ones like the patient's eye or hair colour, may be included in the dataset. The most crucial characteristics, such as blood pressure or cholesterol levels, that are most likely to directly affect the disease's prediction can be found in this situation with the use of feature selection. Since irrelevant factors like hair colour don't affect the forecast, they can be safely eliminated. In order to prevent repetition, one feature may also be eliminated if there is a strong correlation between them (for example, age and weight). By concentrating on the most crucial data, this procedure speeds up, simplifies, and improves the accuracy of the model.

To validate the selected features, we trained predictive models using various classifiers, including Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest. Among these, the Random Forest Classifier consistently delivered the highest accuracy, demonstrating its strength in identifying and leveraging the most relevant features. This thorough selection process ensured that the final feature set optimized model performance while reducing unnecessary complexity.

**3.5.1 Analysis of Variance (ANOVA):**

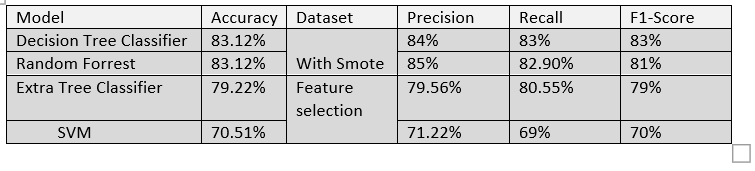
A statistical technique called analysis of variance (ANOVA) is used to examine how the means of three or more groups differ from one another in order to ascertain whether at least one of the groups differs significantly from the others. To put it simply, it assists you in determining whether the variation in your data is attributable to the factor you are evaluating or to pure chance.

**3.5.2 Chi-Square Test:**

The Chi-Square test evaluates the association between target variable and categorical features. It calculates the chi-square statistic for measure the difference between expected distribution and observed. Features with high chi-square values and low p-values are considered the most relevant for cancer prediction.

**3.5.3 Recursive Feature Elimination (RFE):**

Iteratively eliminating the least significant features from a dataset is known as recursive feature elimination, or RFE. It operates by continually constructing a model, assessing each feature's performance to identify which ones add the least to the model's predictions, and then gradually removing those features. Until just the most pertinent features remain for model training, this step is repeated.

****

**Table. 3 Dataset using Smote and Feature Selection**

**3.6 Models Accuracy**

The frequency with which a machine learning model produces accurate predictions is measured by a metric called model accuracy. It is computed by taking the total number of guesses and dividing it by the number of right forecasts. In classification tasks, accuracy is frequently used to evaluate model performance. Accuracy, for instance, indicates how frequently a spam email classifier accurately classifies an email as spam or not. However, when the data is unbalanced—for example, when one class is substantially outnumbered by another—accuracy might be deceptive. In these situations, even if a model performs poorly on the minority class, it may nonetheless appear to be correct when predicting the majority class most of the time. Therefore, in addition to accuracy, it's critical to take into account other metrics like precision, recall, and F1-score.

**3.6.1 Decision Tree Classifier**

One well-liked machine learning approach for classification and regression applications is the Decision Tree Classifier. By dividing the data into subsets according to the feature values, it produces a structure resembling a tree, with each internal node standing for a feature choice and each leaf node for a class label or result. The best feature that maximises the separation of the classes is used to recursively partition the data, which is how the tree is constructed (typically using metrics like Gini impurity or entropy). Decision trees are simple to comprehend, analyse, and illustrate. They are prone to overfitting, nevertheless, particularly when dealing with intricate datasets. Pruning, ensemble approaches (like Random Forests), and setting maximum depth are some ways to avoid this. Despite their ease of use, decision trees can function effectively with

**3.6.2 Random Forest Classifier**

Random Forest is an advanced method that builds multiple decision trees and combines their outputs to improve accuracy and stability. It works by averaging the results from all the trees, reducing the risk of overfitting and enhancing the model's performance. Random Forest is well-suited for complex datasets with many features, making it an excellent choice for analyzing medical data and predicting cancer risks.

**3.6.3 Extra Trees Classifier**

Extra Trees Classifier is similar to Random Forest but introduces additional randomness during the process of building trees. Instead of selecting the best split, it chooses splits at random, leading to a more diverse set of decision trees. This approach often results in better generalization to new data, especially when dealing with noise in the dataset. It is a reliable model for identifying patterns in heart disease data.

**3.6.4 Support Vector Machine (SVM)**

Support Vector Machine is a mathematical model that finds the best boundary to separate different classes of data. By using special techniques (kernels), SVM can create complex decision boundaries, making it effective for handling non-linear relationships in data. SVM is particularly valuable for datasets with many features and helps in accurately identifying individuals at risk of cancers.

**3.7 Proposed Model**

From the start of the training process, the Random Forest model consistently outperformed the other models in terms of accuracy and stability. Without using SMOTE, Random Forest achieved an impressive accuracy of 98.48%, and even after applying SMOTE and feature selection techniques, its performance remained strong at 98.10%. On the other hand, the Decision Tree Classifier also delivered excellent results, with an accuracy of 98.86% after using SMOTE and Chi-Square feature selection. While Decision Tree Classifier is easier to interpret and explain to non-technical stakeholders, it tends to overfit, especially with more complex datasets. This is where Random Forest shines—it offers greater robustness, effectively handling overfitting while maintaining high precision, recall, and F1-score across various datasets.

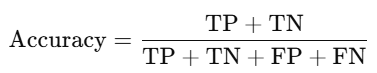
Alongside Random Forest and Decision Tree Classifier, two other models—Extra Tree Classifier and SVM—were also evaluated. Although the Extra Tree Classifier showed decent performance with an accuracy of around 92.42%, it still didn’t match the overall reliability of Random Forest. Meanwhile, SVM’s performance was significantly lower, with accuracy ranging from 65.53% to 68.56%. We also tested an ensemble approach that combined all the models, which yielded the same 98.48% accuracy as Random Forest alone. Despite this, Random Forest was chosen as the final model due to its consistent performance, ability to generalize well to various problems, and robust handling of overfitting. Additionally, Random Forest has the advantage of a relatively lower computational cost compared to the ensemble approach, making it the most efficient and ideal choice for this study.

**3.8 Performance Evaluation Matrices**

A performance evaluation matrix is a logical framework for assessing the effectiveness of a machine learning model. It unfolds the outcomes into accuracy, precision, recall, F1 measure, and the confusion matrix, thus providing a systematic overview of the model’s performance. These various parameters help determine which model is most suitable for a particular problem.

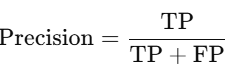
**3.8.1 Accuracy**

It helps us to measure the overall correctness of the model by evaluating the ratio of correctly predicted instances to the total instances.



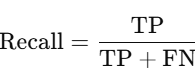
(1)

**3.8.2 Precision**

It evaluates the ratio of correctly predicted positive observations to the total predicted positives.

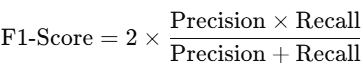
(2)

**3.8.3 Recall**

It helps us to measure the ratio of correctly predicted positive observations to all actual positives.

(3)

**3.8.4 F1-Score**

It represents the harmonic mean of precision and recall, balancing the trade-off between the two.

(4)

**3.8.9 The proper description of TP,TN,FP,FN sequentially in bellow:**

**3.8.9.1 True Positive (TP):** True positive means when the model correctly predicts a cancer, and the person actually has a cancer.

**3.8.9.2 True Negative (TN):** Meaning of true negative is when the model correctly predicts no cancer, and also the person doesn’t have any cancer.

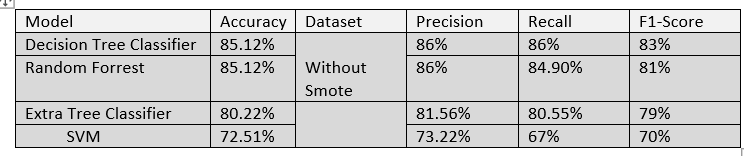
**3.8.9.3 False Positive (FP):** This is when the model incorrectly predicts a cancer, but the person does not actually have a cancer.

**3.8.9.False Negative (FN):** Last one false negative which means when the model incorrectly predicts no cancer, but the person actually has a cancer.

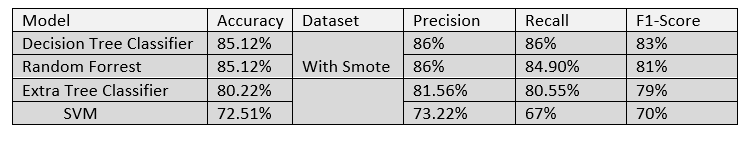
**3.9 Result**

We assessed the performance of four machine learning models Random Forest (RF), Decision Tree (DT), SVM, and Extra Trees Classifier for predicting cancer risk. we used key metrics such as accuracy, precision, recall, and F1-score to evaluate how well each model identifies cancer disease. We applied smote and feature selection to enhance the models performance. The showed how each algorithm performs, offers useful insights into their potential for accurately predicting cancer risk in clinical settings.

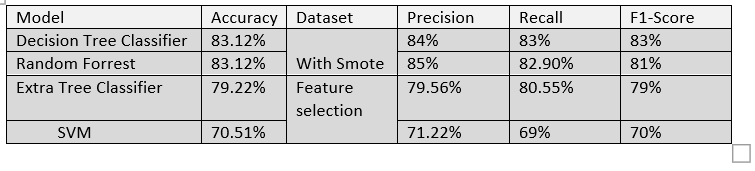
For the normal dataset we got highest accuracy of 98.48% using Random Forest (RF), showing strong performance in predicting cancer disease and consistently well performed with others in key metrics like precision, recall, and F1-score, making it the most dependable. While the Decision Tree (DT) model achieved a close accuracy of 97.72%, it was prone to overfitting, which made it less reliable. The Extra Trees Classifier (ETC) showed a respectable accuracy of 92.42%, but it couldn’t match the overall performance of RF and DT. Meanwhile, the Support Vector Machine (SVM) had the lowest accuracy, ranging from 65.53% to 68.56%, indicating it wasn’t as effective for cancer prediction

**Table. 1 Dataset without using Smote**

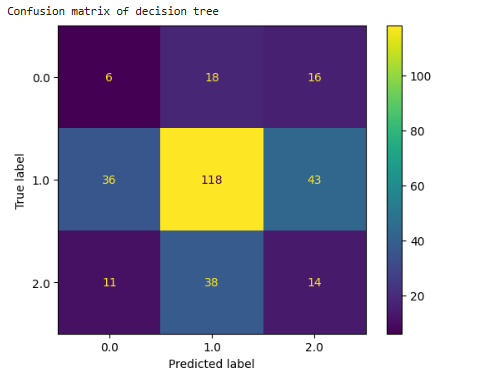
After applying SMOTE to the dataset, Random Forest (RF) showed strong performance with an accuracy of 98.10% maintaining its position as the most reliable model, robustness and performing consistently well across key metrics such as precision, recall, and F1-score. Decision Tree (DT) model achieved similar accuracy as RF after SMOTE. The Extra Trees Classifier (ETC) performed consistently with an accuracy of 92.42%, but it remained less effective than RF and DT. On the other hand, the Support Vector Machine (SVM) showed a slight improvement in accuracy, ranging from 66.28% to 68.56%, but it still lagged behind the other models, making it less effective for cancer prediction when compared to RF and DT after SMOTE balancing.

**Table. 2 Dataset using Smote**

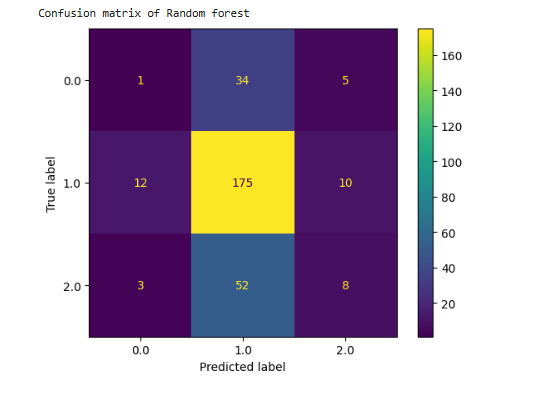
After applying feature selection techniques in the dataset with smote, Random Forest (RF) continued to shine with an accuracy of 98.10%, reaffirming its position as the most reliable model for predicting cancer risk. It performed consistently well across key metrics like precision, recall, and F1-score, proving its robustness even after the additional steps of SMOTE and feature selection. The Decision Tree (DT) model also reached an accuracy of 98.10%, but it still showed signs of overfitting, which made it a bit less reliable than RF in practical use. The Extra Trees Classifier (ETC) maintained an accuracy of 92.42%, but it didn’t quite measure up to RF and DT in terms of overall performance. The Support Vector Machine (SVM) saw only a slight improvement in accuracy, ranging from 66.28% to 68.56%, and still lagged behind the other models, making it less suitable for cancer prediction even after SMOTE and feature selection.

**Table. 3 Dataset using Smote and Feature Selection**

In the below given figure, in the case of a normal dataset DT has 154 TP, 4 FP, 2 FN and 104 TN values. In the case of the DT after applying smote 155 TP, 3 FP, 1 FN and 105 TN. IN case of ANOVA 156 TP, 2 FP, 2 FN and 104 TN. IN Case of Chi-Square 157 TP, 1 FP, 2 FN and 104 TN and at last for DT RFE 154 TP, 4 FP, 2 FN and 104 TN.

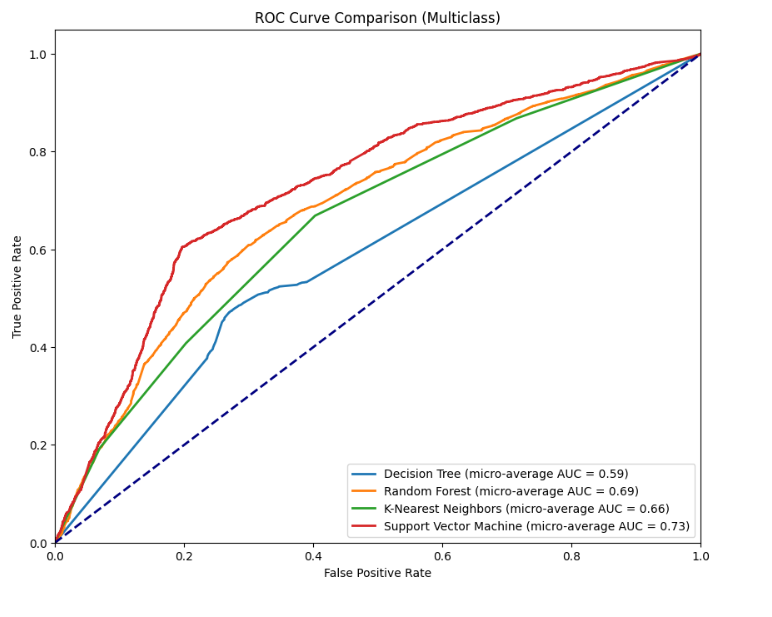


In the below given figure, in the case of a normal dataset RF has 154 TP, 4 FP, 2 FN and 104 TN values. In the case of the RF after applying smote 154 TP, 4 FP, 1 FN and 105 TN. IN case of ANOVA 154 TP, 4 FP, 2 FN and 104 TN. IN Case of Chi-Square 155 TP, 3 FP, 2 FN and 104 TN and at last for RF RFE 154 TP, 4 FP, 1 FN and 105 TN.In the below given figure, in the case of a normal dataset SVM has 153 TP, 5 FP, 86 FN and 20 TN values.



In the case of the SVM after applying smote 92 TP, 66 FP, 27 FN and 79 TN. IN case of ANOVA 147 TP, 11 FP, 72 FN and 34 TN. IN Case of Chi-Square 147 TP, 11 FP, 72 FN and 34 TN and at last for SVM RFE 149 TP, 9 FP, 74 FN and 32 TN.

In the below given figure, in the case of a normal dataset ETC has 146 TP, 12 FP, 8 FN and 98 TN values. In the case of the ETC after applying smote 140 TP, 18 FP, 11 FN and 95 TN. IN case of ANOVA 153 TP, 5 FP, 4 FN and 102 TN. IN Case of Chi-Square 152 TP, 6 FP, 4 FN and 102 TN and at last for ETC RFE 151 TP, 7 FP, 3 FN and 103 TN.



**Fig. 10 Receiver operating characteristic (ROC) curve.**

**4. Conclusion**

Based on input features, a classifier is a kind of machine learning algorithm that groups data into distinct classes or labels. It learns from a labelled dataset, finds links and patterns in the data, and then uses these patterns to predict the class of data that isn't visible. K-Nearest Neighbours, Naive Bayes, Decision Trees, and Support Vector Machines are examples of popular classifiers. Classifier performance is assessed using metrics such as F1-score, recall, accuracy, and precision. Classifiers are strong tools for tasks like picture identification, medical diagnosis, and spam detection, but how well they work relies on the data's quality, the algorithm used, and how it is tuned. Effective assessment and validation are essential to guaranteeing that the classifier performs well when applied to novel, unseen data.

R**eference**

1. [1] D. Crosby et al., “Early Detection of Cancer,” Science, vol. 375, no. 6586, Mar. 2022, doi:<https://doi.org/10.1126/science.aay9040>.
2. [2] N. M. ud din, R. A. Dar, M. Rasool, and A. Assad, “Breast cancer detection using deep learning: Datasets, methods, and challenges ahead,” Computers in Biology and Medicine, vol. 149, p. 106073, Oct. 2022, doi:<https://doi.org/10.1016/j.compbiomed.2022.106073>.
3. [3] M. Tahmooresi, A. Afshar, B. B. Rad, K. B. Nowshath, and M. A. Bamiah, “Early Detection of Breast Cancer Using Machine Learning Techniques,” Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 10, no. 3–2, pp. 21–27, Sep. 2018, Available:<https://jtec.utem.edu.my/jtec/article/view/4706>
4. [4] A. E. Pelzer et al., “PROSTATE CANCER DETECTION IN MEN WITH PROSTATE SPECIFIC ANTIGEN 4 TO 10 NG/ML USING A COMBINED APPROACH OF CONTRAST ENHANCED COLOR DOPPLER TARGETED AND SYSTEMATIC BIOPSY,” vol. 173, no. 6, pp. 1926–1929, Jun. 2005, doi:<https://doi.org/10.1097/01.ju.0000158444.56199.03>.
5. [5]F. Aigner et al., “Value of Real-Time Elastography Targeted Biopsy for Prostate Cancer Detection in Men With Prostate Specific Antigen 1.25 ng/ml or Greater and 4.00 ng/ml or Less,” vol. 184, no. 3, pp. 913–917, Sep. 2010, doi:<https://doi.org/10.1016/j.juro.2010.05.026>.
7. [6] A. F. M. Agarap, “On breast cancer detection,” Proceedings of the 2nd International Conference on Machine Learning and Soft Computing - ICMLSC ’18, 2018, doi:<https://doi.org/10.1145/3184066.3184080>.
8. [7] T. Tamada et al., “Prostate Cancer Detection in Patients With Total Serum Prostate-Specific Antigen Levels of 4–10 ng/mL: Diagnostic Efficacy of Diffusion-Weighted Imaging, Dynamic Contrast-Enhanced MRI, and T2-Weighted Imaging,” American Journal of Roentgenology, vol. 197, no. 3, pp. 664–670, Sep. 2011, doi:<https://doi.org/10.2214/ajr.10.5923>.
9. [8] Xu, C.-X. (2014). Diabetes and cancer: Associations, mechanisms, and implications for medical practice. World Journal of Diabetes, 5(3), 372.<https://doi.org/10.4239/wjd.v5.i3.372>
10. [9] Miller, M., Mogun, H., Azrael, D., Hempstead, K., & Solomon, D. H. (2008). Cancer and the Risk of Suicide in Older Americans. Journal of Clinical Oncology, 26(29), 4720–4724.<https://doi.org/10.1200/jco.2007.14.3990>
11. [10] Fortuny, J., Sima, C., Bayuga, S., Wilcox, H., Pulick, K., Faulkner, S., Zauber, A. G., & Olson, S. H. (2009). Risk of Endometrial Cancer in Relation to Medical Conditions and Medication Use. Cancer Epidemiology, Biomarkers & Prevention, 18(5), 1448–1456.<https://doi.org/10.1158/1055-9965.epi-08-0936>
12. [11] Schultz, P. N., Beck, M. L., Stava, C., & Vassilopoulou-Sellin, R. (2003). Health profiles in 5836 long-term cancer survivors. International Journal of Cancer, 104(4), 488–495.<https://doi.org/10.1002/ijc.10981>
13. [12] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning. 2018 Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT).<https://doi.org/10.1109/ebbt.2018.8391453>
14. [13] Cho, H., Mariotto, A. B., Mann, B. S., Klabunde, C. N., & Feuer, E. J. (2013). Assessing Non–Cancer-Related Health Status of US Cancer Patients: Other-Cause Survival and Comorbidity Prevalence. American Journal of Epidemiology, 178(3), 339–349. https://doi.org/10.1093/aje/kws580
15. [14] Yabroff, K. R., Lawrence, W. F., Clauser, S., Davis, W. W., & Brown, M. L. (2004). Burden of illness in cancer survivors: findings from a population-based national sample. Journal of the National Cancer Institute, 96(17), 1322–1330.<https://doi.org/10.1093/jnci/djh255>
16. [15] Brignardello, E., Felicetti, F., Castiglione, A., Chiabotto, P., Corrias, A., Fagioli, F., Ciccone, G., & Boccuzzi, G. (2013). Endocrine health conditions in adult survivors of childhood cancer: the need for specialized adult-focused follow-up clinics. *European Journal of Endocrinology*, *168*(3), 465–472.<https://doi.org/10.1530/eje-12-1043>
17. [16] Welch, H. G., & Black, W. C. (2010). Overdiagnosis in Cancer. JNCI Journal of the National Cancer Institute, 102(9), 605–613.<https://doi.org/10.1093/jnci/djq099>
18. [17] Cestari, D. M., Weine, D. M., Panageas, K. S., Segal, A. Z., & DeAngelis, L. M. (2004). Stroke in patients with cancer: Incidence and etiology. Neurology, 62(11), 2025–2030.<https://doi.org/10.1212/01.wnl.0000129912.56486.2b>
19. [18] Hewitt, M., Rowland, J. H., & Yancik, R. (2003). Cancer survivors in the United States: age, health, and disability. The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences, 58(1), 82–91.<https://doi.org/10.1093/gerona/58.1.m82>
20. D. Jalal and Adnan Mohsin Abdulazeez, “A Review on Heart Disease Detection Classification Based on Deep Learning Algorithm,” Indonesian Journal of Computer Science, vol. 13, no. 2, Apr. 2024, doi: <https://doi.org/10.33022/ijcs.v13i2.3921>.
21. ‌A. Verma, G. Agarwal, A. K. Gupta, and M. Sain, “Novel Hybrid Intelligent Secure Cloud Internet of Things Based Disease Prediction and Diagnosis,” Electronics, vol. 10, no. 23, p. 3013, Dec. 2021, doi: <https://doi.org/10.3390/electronics10233013>.
22. ‌A. Choi et al., “Development of Prediction Models for Acute Myocardial Infarction at Prehospital Stage with Machine Learning Based on a Nationwide Database,” Journal of Cardiovascular Development and Disease, vol. 9, no. 12, pp. 430–430, Dec. 2022, doi: <https://doi.org/10.3390/jcdd9120430>.
23. ‌I. A. Kakadiaris, M. Vrigkas, A. A. Yen, T. Kuznetsova, M. Budoff, and M. Naghavi, “Machine Learning Outperforms ACC/AHA CVD Risk Calculator in MESA,” Journal of the American Heart Association, vol. 7, no. 22, Nov. 2018, doi: <https://doi.org/10.1161/jaha.118.009476>.
24. [Izabela Rojek](https://link.springer.com/chapter/10.1007/978-3-642-13232-2_12#auth-Izabela-Rojek) Hybrid Neural Networks as Prediction Models
25. I. Rojek, M. Kozielski, J. Dorożyński, and D. Mikołajewski, “AI-Based Prediction of Myocardial Infarction Risk as an Element of Preventive Medicine,” Applied Sciences (2076-3417), vol. 12, no. 19, p. 9596, Oct. 2022, doi: <https://doi.org/10.3390/app12199596>.
26. “Data Splitting in Machine Learning,” *CloudFactory Computer Vision Wiki*. <https://wiki.cloudfactory.com/docs/mp-wiki/splits/data-splitting-in-machine-learning>
27. C. GOYAL, “How to Detect and Remove Outliers | Outlier Detection And Removal,” *Analytics Vidhya*, May 19, 2021. <https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/>
28. H. Bonthu, “Detecting and Treating Outliers | How to Handle Outliers,” *Analytics Vidhya*, 2021. <https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/>
29. S. Mazumder, “What is Imbalanced Data | Techniques to Handle Imbalanced Data,” *Analytics Vidhya*, Jun. 21, 2021. <https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/>
30. “7 SMOTE Variations for Oversampling,” *KDnuggets*. <https://www.kdnuggets.com/2023/01/7-smote-variations-oversampling.html>
31. A. Gupta, “Feature Selection Techniques in Machine Learning,” *Analytics Vidhya*, Oct. 10, 2020. <https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/>