European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 07, Issue 11, 2020

Applying an Optimal Feature Ranking and Selection Algorithm and Random Forest Classifier Algorithm along with K-Fold Cross validation for Classification of Blood Cancer Cells

Mukesh Madanan[1], Anita Venugopal[2] and Nitha C.Velayudhan[3]

[1] Dhofar University, Salalah,Postal Code 211,Sultanate of Oman
[2] Dhofar University, Salalah,Postal Code 211,Sultanate of Oman
[3] Noorul Islam Centre for Higher Education, Tamil Nadu, India

ABSTRACT

Technological inventions and researches have pushed various emerging fields to overwhelm information technology and its applications. Bioinformatics, medicines and drug discovery have found significance use of Artificial intelligence for aiding in various discoveries and patient centered care. One of the most critical area of medical concern is blood cancer that has limitations in detection and classification at a proper stage. By applying artificial intelligence techniques to the blood genetics, physicians and oncologists all over the world could find a solution for early detection and classification of the disease. Susceptibility, recurrence and survival rates of cancer are questions to be answered while detecting and classifying stages of cancer. Artificial intelligence techniques and models proposed by researchers for characterizing various stages of blood cancer have aided in providing treatment to certain extend. Applying artificial intelligence framework models on genetics has been found to provide results that satisfies the detection of the disease. However, an efficient methodology that focuses of selecting particular features of blood cancer and ranking these features and then classifying them into various stages is still lacking. This research focuses of using an optimal feature selection and ranking algorithm for ranking and selecting features and then employing random forest algorithm for classifying the blood cancer stages in the blood sample. This technique allows to provide better patient centered care and early classification of cancer stages.

Keywords: *Artificial Intelligence, Machine Learning Algorithm, Deep Learning, Feature Selection, Random Forest Classifier, Blood Cancer Detection*

1 Introduction

Machine learning and deep learning have paved its way into healthcare for diagnosing and predicting diseases to certain extend. Numerous decision support systems to aid clinical experts in decision-making have been created using machine learning algorithms. Utilizing these systems, the physicians' diagnose various ailments and could easily provide treatment and medication for the patients. In the absence of these advanced systems, the physicians does preliminary check-ups on the patient and based on various tests done make a diagnosis and then using their knowledge prescribe treatment and medicines. In such cases, the question arises on the experience and knowledge of the clinical expert as no solid data is available on the accuracy of the judgement based on check-ups and test results. Moreover, the medical history of the patient also holds a vital key in the diagnosis and treatment. Availability of the entire medical history and health records of the patient is also not feasible during emergencies. Digital and electronic medical records of patients have given rise to the availability of these resources at the physician's disposal. Linking the decision support system to the medical records of patient could make it easier for the physician to access the medical history and certainly provide much more accurate diagnosis and prescribe treatment. The decision support system keeps track of such medical records and the treatment provided which in turn can be used in future as a reference and can in fact predict the treatment for similar cases in future. This process is carried out by the decision support system by classifying various ailments and then mapping each aliment to a treatment. The classification tasks and mapping are the highlights of the decision support systems. All the classification and mapping

is based on correlated properties of the relevant data of the medical history of the patient that is the major obstacle of the conventional technologies.

Cancer is an ailment that results from a group of diseases [1].It is not a standalone disease that inhibits the body [1]. IN a normal human body, older cells die and are replaced by new cells as per Al-Shamasneh and Obaidellah [2]. In certain situations in the human body some adverse effects takes place resulting in the extended lifetime of the older cells and these older cells begin to grow. This abnormality results in the survival of these older cells which in turn leads to stage called cancer. Oncologists classify cancer as Leukaemia, Lymphoma and Myeloma. Leukaemia is a condition associated with white blood cells of the human body. In the human body, white blood cells are considered the infection fighters and they grow normally and multiply when an infection occurs in the body. Leukaemia is a condition in which the bone marrow produces abnormal white blood cells leading to a situation that impairs the red blood cell and blood platelets development. Lymphoma is a condition which effects the lymphocytes. Lymphocytes are the infection-fighting cells of the immune system located in the lymph nodes, spleen, bone marrow etc. When lymphoma occurs these cells grow out of control. In myeloma, antibody production is impaired as it effects the blood plasma. This situation results in a case of very weak immune system. When cancer occurs, depends on the cells growths the body shows various health issues as it has divergent effects in patient to patient. When the cancerous cells become dominant, the body caves in this situation as the immunity is reduced. This happens at the penultimate stage of cancer.

Artificial intelligence methods have provided new ways for predicting the occurrence and detecting the cancerous cells in human body. The most important role of artificial intelligence is in solving complex tasks as per Pavandeh et al [14]. Machine learning and Deep learning have found its applications in various health providing systems to solve complex issues. Due to this, researchers and scientists have started using artificial intelligence for image processing of cancer cells and this in turn aids in timely diagnosis. Early diagnosis can certainly speed up treatment and thus increase survival rates in patients.

The research grails to design a methodology for the classification of blood cancerous cells based on the National Center for Biotechnology Gene Expression Omnibus database. The database contains various medical data regarding the blood cancer cells but it is considered a limited dataset. Extreme Learning Machine (ELM) is an example for similar approach. Estay, Faris and Obeid proposed an ELM based competitive Swarm Optimisation approach [4].The research paper focuses on developing a methodology model for feature selection and ranking the features of blood cancer cells initially. The selected features after ranking are then classified into various stages of blood cancer. The feature selection and classification algorithm are the basis of the prediction model. The results obtained could be employed for early detection and classification of blood cancer and diagnosing and treatment thus providing patient centered care. Section 2 focuses on how artificial intelligence can be used and how it is employed in blood cancer detection.

2 Literature Review

Artificial methods have been used significantly in various fields for predicting results such as in agriculture, finance, stock market, weather forecasting etc. Image processing, speech and voice applications have all utilized artificial intelligent methods. Recently Medicine and Drug discovery have also found use of artificial intelligence and its sub fields such as machine learning and deep learning. Many researchers have also applied artificial intelligence techniques in predicting blood cancer.

Enhancing human life style is what technology has done in this era. This has not only provided facilities at ones disposal but also provided a better understanding of how technical innovations in medical field could save lives. Healthcare has been blessed with the inventions where artificial intelligence methods could be used to make predictions about occurrence of blood cancer and then detecting it and providing

treatment. Digital technology has allowed assemblage and compilation of comprehensive genetic data affiliated to blood cancer. Clinical experts, physicians and oncologist use this archived genetic data various research for ailment detection and drug or vaccine discovery [2]. Advanced genetic data information could be used for prediction of the occurrence of blood cancer too. Physicians and oncologists often find it difficult to make accurate predictions regarding the blood cancer and it is considered a major challenge. Image processing and pattern recognition are employing machine learning algorithms to better comprehend the relationships from various data sets in order to make accurate and precise predictions for blood cancer occurrence and detection[6]. Researchers have proposed various models for predicting susceptibility of patients to blood cancer. Several other models have also been proposed to predict the recurrence and survival rates in patients. Due to the flexibility and adaptability of these artificial intelligence techniques and machine learning algorithms they have become the mainspring of susceptibility prediction of blood cancer. Profuse models are also at hand which can predict the cancer stages.

2.1 Predicting Blood Cancer with Artificial Intelligence Models

The primary aim of blood cancer stage classification is to categorise the level of risk the patient might be in at the time of disease diagnosis. The technical innovations made by artificial intelligent algorithms have paved way for the development of efficient decision support systems which can predict the stage and therapy for blood cancer patients. The process will involve the analysis of cells and then classifying them. However, a thorough validation process is essential for the existing models to be considered effective. These validation levels must be approved so that the clinical experts can blindly follow the classified data and prescribe treatment for the patients [2]. These models could be could be used for susceptibility and recurrence prediction of blood cancer in patients.

2.1.1. Susceptibility Prediction Models for Blood Cancer

One of the most used algorithms for predicting the susceptibility of blood cancer in patients is the Artificial Neural Networks(ANN). The feature of the model is the existence of numerous identical layers which are responsible for generalisation. In order to make prediction data analysis is performed on multiple records and data sets and then these data are fed into the model. This is done to enable the model to distinguish between cancerous cells and normal blood cells. Radiologists perform evaluation and revision of these data before it is fed into the model. A typical structure of Artificial Neural Network with multiple hidden layer is depicted in Fig 1.

Predicting the susceptibility of the cancerous blood cells in patients is facilitated by the use of Machine Learning algorithms. Learning Classification systems are the most utilised technique in this aspect. The classification of the malignant and benign blood cells could be easily be performed by the prediction model [7]. The model also finds its use in the prediction of breast cancer and hence it could also be employed in blood cancer prediction. One unique feature is that the method utilises big data to depict the cancer ballooning and spread in the blood. In order to discriminate between genial and malevolent blood cells , the machine learning applies the discrimination technique. The most notable feature of this technique and method using ANN is in the early stages where it has proven to detect blood cancer much effectively. This is due to the detection it can make in the augmentation of the cells in the liver, lymph nodes and the spleen.

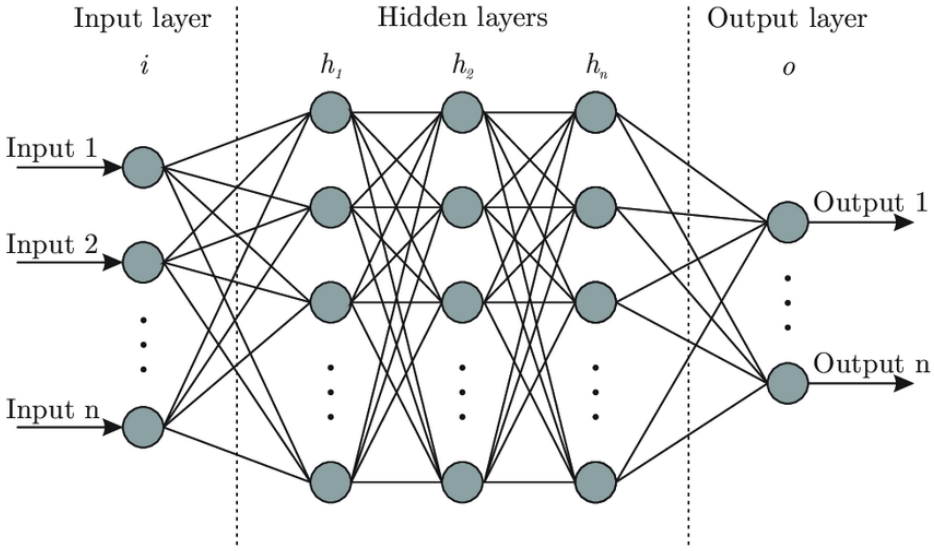

Fig 1. An illustration of the ANN Structure

2.1.2. Reoccurrence Prediction Models for Blood Cancer

The reoccurrence of the blood cancer in many patients have raised doubts about the treatment given to patients even if they recover once. To predict the reoccurrence of blood cancer, machine-learning models could be conceptualized. The prediction of reoccurrence is much essential for patients who are survivors of the disease as this can certainly aid in continuing medication and treatment. Physicians and oncologists often find it arduous to arbitrate whether the survivors could alight upon the cancer again or not. Prediction of reoccurrence has attained success to certain extent due to the application of artificial intelligence[7]. A series of machine learning algorithms and deep learning methods have successfully been able to predict whether a person could be blood cancer patient again. For the prediction, a lot of clinical data, genomic information and cell images are collected and provided as input to the models. Commonly for the development of such models, support vector machines(SVM), Decision Trees Narrow Down to an Outcome (DT model) and the Bayesian Networks Estimate Probability(BN) are utilised [7].In addition to these models, Oral Squamous cell Carcinoma (OSCC) is also employed. For analysing the data, it is fed as input to the Decision Support System and later the OSCC can be integrated to make predictions regarding the transmission factor [5].For making an efficient prediction lot of data is required and fed into the system. To support the prediction process, a wrapper algorithm enforced [7].

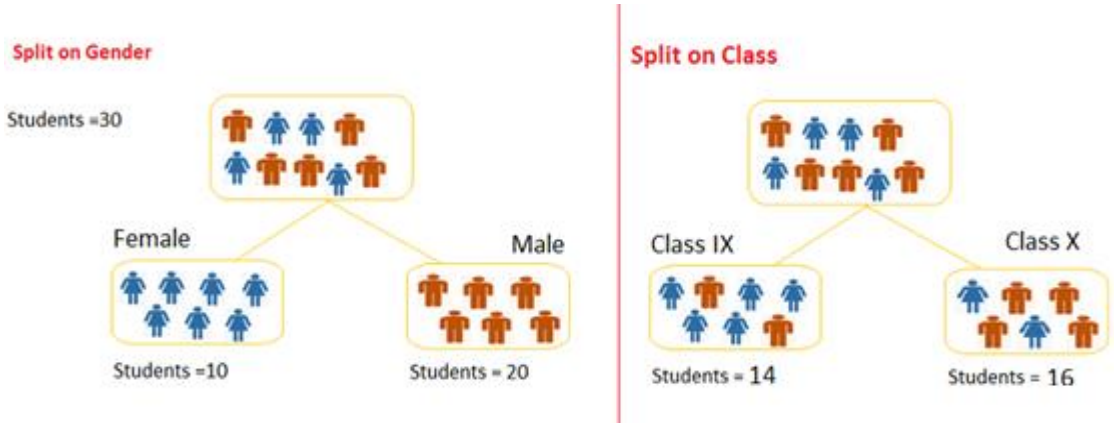

Fig 2. Decision Tree Split on Gender and Class

The prediction process enforces the detection of blood cancer cells in patients who have already survived and then discriminates the patients into groups of positive and negative cancer cells. In this

method, Bayesian Networks, Artificial neural Networks, Support vector machines and even Random Forest algorithms are selected for the process of classification of blood cells [7]. Following this process, a validation such as 10-fold cross-validation is used to evaluate the whole process. Receiver operating characteristic curve (ROC) is inked in to generalise the data. This method has been successful and has provided positive results in the prediction of reoccurrence of blood cancer and other types of cancer. The Fig 2 and Fig 3 depicts a Decision Tree hinging on machine learning.

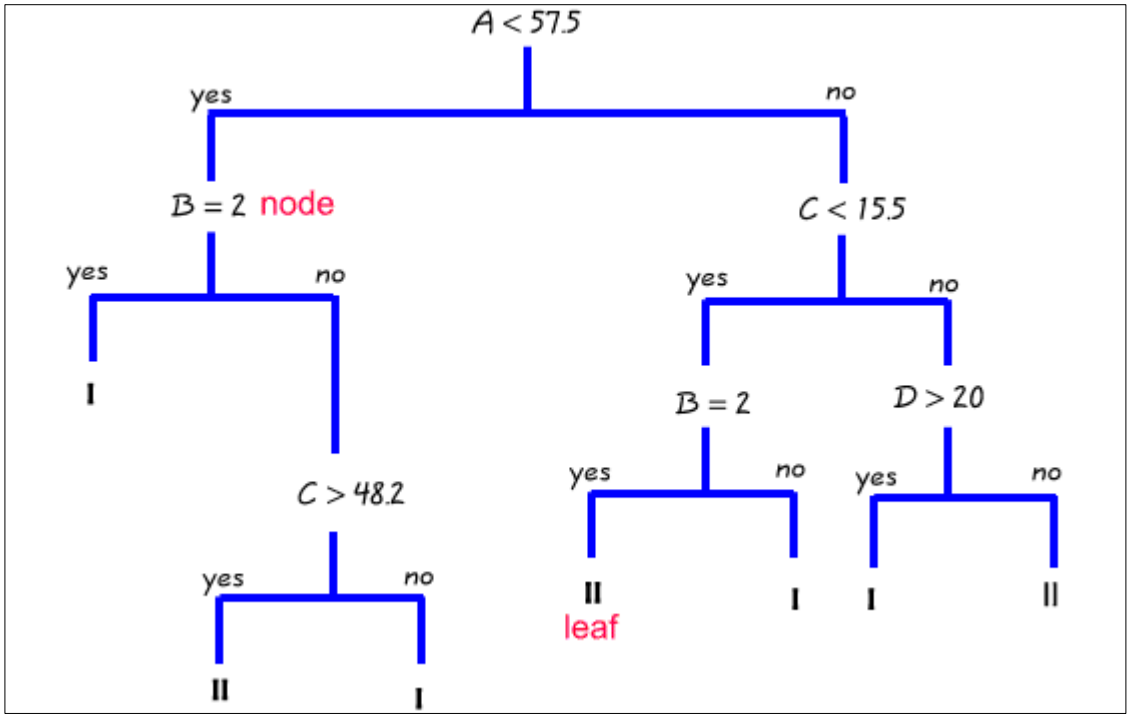

Fig 3.Decision Tree Depicting Classifying to Class I and Class II

2.1.3. Survival Rate Prediction Models for Blood Cancer

When a patient is diagnosed with cancer, his/her chance of survival is an important factor. The survival rates depends on numerous factors including the stage at which the cancer was detected, the patients' medical history, patients' response to treatment and overall the causes of the cancer. Physicians and oncologists recommend lot of tests and treatment but they are also in dilemma to know the survival rate of these patients. Researchers have performed numerous research and tests ad have been successful to determine survival rates of the patients using artificial intelligence techniques which has provided significant evidence and reliability.

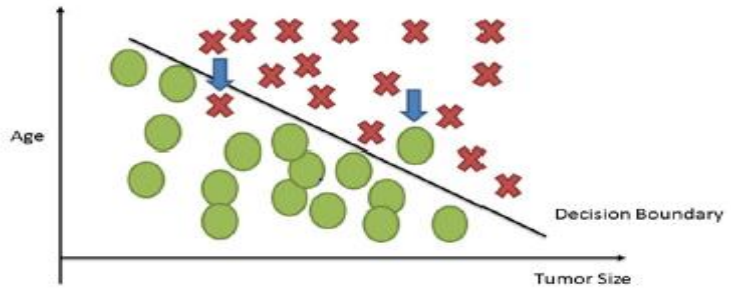

Fig 4. The SVM classification Technique [7]

Research conducted by Kourou and et al, provided compelling testimony regarding the survival rates of patients using artificial intelligence techniques [7].The patients were classified in their study based on the cancerous cell size and their age [7]. The model was put to test using Support Vector Machines,

Artificial Neural Networks and Semi-supervised Learning that provided ample repercussion with a 65%, 51% and 71% success [7].The focus of their research was the size of the tumour along with the time of detection [7].A fivefold cross validation was used to validate the classification [17]. Another factor which was considered was the quantity of nodes [7]. Based on this classification the survival rates could be predicted as size of cells is an indication of the growth. This is major linchpin in the research to detect the survival rates of blood cancer patients [7]. The question of survival of patients is mainly focused if the disease is detected in the penultimate or last stage. Hence, the model is applied usually in the last stage.

2.1.4. Alternative Models employed

Voluminous researches have been conducted in order to predict cancer. Molecule by molecule of the cancer cells have been targeted and application of the Quantgene technology have flourishingly set the ball rolling to detect blood cancer in the early stages. Quantgene technology fixates on the DNA centric technique [15]. Screening and Genomics both detect cancerous cells in this method. The cell-free DNA is analysed in this method and this data is then used for assaying and determining the genetic complexion of the cancerous cells. It also provides cue regarding the location of the cancer. The method involves segregation of the DNA cell copies in the blood stream [15]. In fact, cfDNA cells are mutated and this distinguishes the cancerous cells from genetic DNA. Pattern recognition is complex as several mutation patterns relationships are involved when compared to neural networks in the detection of cancerous cells.

University Health Network in Canada implied the adoption of artificial intelligence methods to detection and classification of cancer at the antecedent stages [12]. The techniques harnessed include liquid biopsy and epigenetic amendments. After these, it is applied to machine learning algorithms to perform tests. Additionally, to detect cancer attached DNA could be detected using big data analysis.

3 Machine Learning

To detect and classify cancerous cells, artificial intelligence utilises machine-learning algorithms. Cancer forecasting and visualisations could be easily be attained by machine learning algorithms. Depending on contingent probabilities and prohibitive alternatives machine learning algorithms does calculations and accomplishes the classification process. Most proficient Machine Learning algorithms are Artificial Neural Networks and Decision trees.

3.1 Artificial Neural Networks

According to Kourou et al, artificial neural networks are convenient for handling a pattern recognition and classification tasks at a very large scale [7]. ANN features demonstrate in having the preference to handle a scope of measurable, such as nonlinear, linear, and logistic regression well balanced with logical tasks or derivations, such as AND, OR, XOR, NOT, and IF-THEN as a feature of the classification and recognition strategies. ANNs are depiction of how neurons in the brain communicate with each other through axon intersections. Kourou et al states that additionally alongside organic learning the quality of the neural associations is barricaded or weakened through extended preparing or fortification on marked training data and information [7].

A wiring table or framework represents the neural associations in ANNs. Precisely it can be depicted as neuron 1 is connected to neuron 2, 4 and 7 and neuron 2 is associated with neuron 1,5,6 and 8 and so on [16]. Layers it he term given to this weight grid and is very much in similar to the cortical layers in the brain cerebrum. Concealed layers are ones which are utilised by the neural systems to process data and information and produce a decision as depicted in Fig 5. According to Manogaran et al, in ANNs

to represent input and output, numbers, vectors and strings are used as numerical structure of each layer [10].

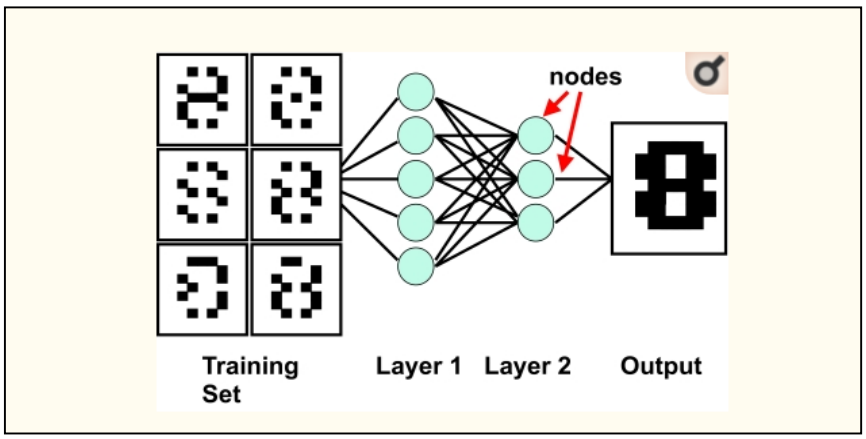

Fig 5: Machine learning training technique to identify images through a training set [10]

3.2 Decision Trees

For cancer detection and diagnosis, decision trees can be a comprehensive method that provides reliability. According to Liu et al. decision trees are coordinated, correlated diagrams streamlined with numerous options for selection. The nodes in the tree along with the possible and probable decisions (leaves or branches) are arranged to deduce an objective [8].Decision trees are particularly utilised in categorization of scientific data and are considered to be an indispensable part of medical conventions [8]. A decision tree showing the breast cancer diagnosis is shown in Fig 6.

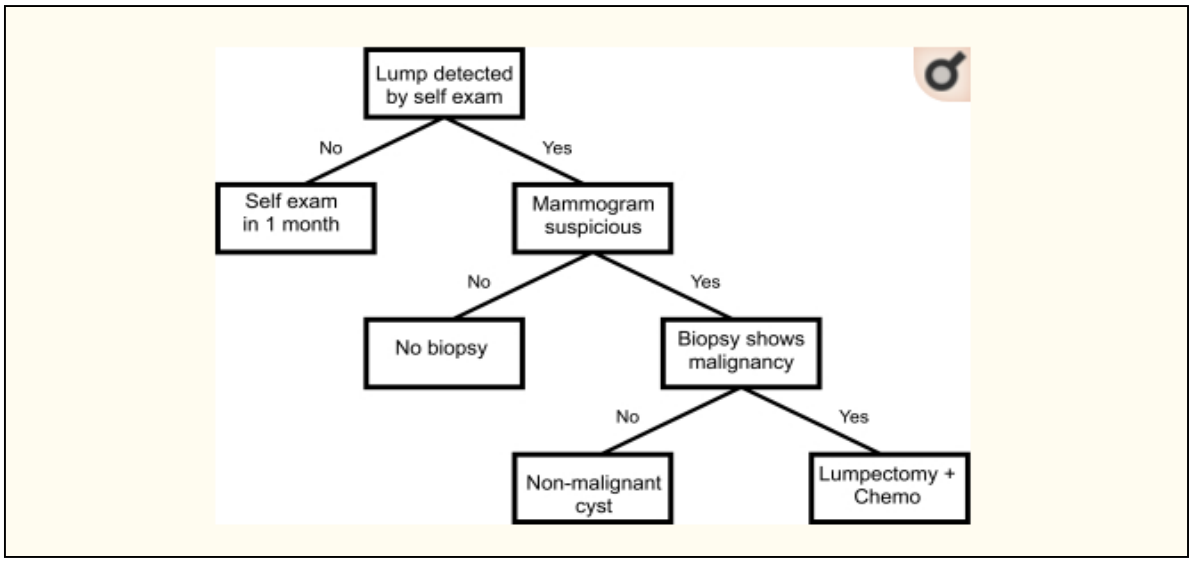

Fig 6: Decision tree used in diagnosis and treatment of breast cancer [8]

Consultations and interviews are used for the creation of decision trees so that a conclusive objective could be attained. The personals involved in the creation and discussions are the specialists in the area. The decision trees are then drafted through long discussions and agreements. Sometimes the decision trees are also alterations of the existing ones to consent to asset restrictions or to confine chance. However, on giving a named set of trained data, the decision trees can themselves create learners. When the learners for decision trees are utilised to distinguish data, the ‘leaves’ in the decision tree indicate classifications, whereas the ‘branches’ denote the associations that forefront to those characterizations. The training data are segregated frequently into datasets. These datasets are then replenished to decision trees for learning process depending on either logical or numerical tests. Thereupon, this procedure is

summarised and modified on particular resolute subset in a repeated means until partitioning is unessential or when a lonely classification is achieved [3]. Decision trees have bountiful assenting prospects in cancer revealing and diagnosis: exemplifying, they are offhand to comprehend and translate, they need meagre information planning, they can work with copious assortment of information including numeric, ostensible and all out information, they yield powerful classifiers, they 'learn' instantaneously, and they are endorsed using factual tests.

4 Deep Learning

In numerous cases, such stroke injury segmentation and Alzheimer analysis we require a thorough image examination of the cerebrum. The artificial intelligence method appropriate for image analysis is deep learning. The most prominent deep learning algorithm employed for image analysis is the Convolutional Neural Network (CNN). CNN is best used for classifying and segmenting images of the human body in medical diagnosis.

4.1 Convolutional Neural Network(CNN)

For cancer imaging, convolutional neural networks identifies spatial connection between pixels. This process is done in a hierarchical manner. According to Vogado et al, medical imaging of the human body is done by convolving medical images. This process of convolution is done through scholarly channels which in turn create a chain of mapping highlights [18]. The process of convolution s carried out in numerous layers and this determines the capacity of convolution. The highlight achieved through the process are the interpretations that could be the high level of exactness. The CNN has several layers which are discussed as follows:

a) Input Image Format Layer:

A variety of pixels values determine the image input. These pixels depend on the size of the picture according to Lopez-Rincon et al [9]. For instance a 3 × m × n cluster of numbers depicts a shaded information picture. Here '3' signifies the values blue, green, and red with the values of the pixels for each shading running from 0–255; also, m and n are the elements of the picture[9]. On account of a grayscale picture, the picture size is characterized by 2D exhibit (m × n), where the force of the pixels likewise extends from 0–255 [9].

b) Convolution Layer:

Convolutional Layer is the most prominent and important layer of CNN. This layer using the convolutional channels captures highlights from the provided input image. These channels are square exhibits with numbers which characterise loads and parameters. The top left corner of the image is actually the initial position of the channel in the convolutional process. In this process, the duplication of picture pixel lattice progressively. In addition, the channel grid is also progressed. According to Mocan et al, adding the pixel lattice and channel grid rehashes the sliding channels to one [11].The strides indicate the amount of cell movements to one side during each and every progression [11].

5 Proposed Design and Methodology for Classifying Blood Cancer Cells

The research proposes a smart method to facilitate the detection and classification of blood cancer. By doing so physicians and oncologists can provide proper timely treatment to the patients. This can actually increase the survival rates of patients. Fig 7 depicts the proposed architecture. The proposed method includes initial data acquisition process followed by data pre-processing step. From the pre-processed data, a data set is created for training aspect. Subsequently, certain features from the data set are then ranked using machine learning algorithms. After ranking, features are selected and then it is applied to a feature classifying algorithm. Finally, a prediction model is created based on a classification algorithm. The architecture has the following modules:

(i) Data Acquisition
(ii) Data Pre-processing
(iii) Dataset Creation
(iv) Ranking of Features of Datasets
(v) Selection of Ranked Features
(vi) Removal of unranked Features
(vii) Application of Classification Algorithm to the selected ranked features
(viii) Construction of a Blood Cancer stage Classifier

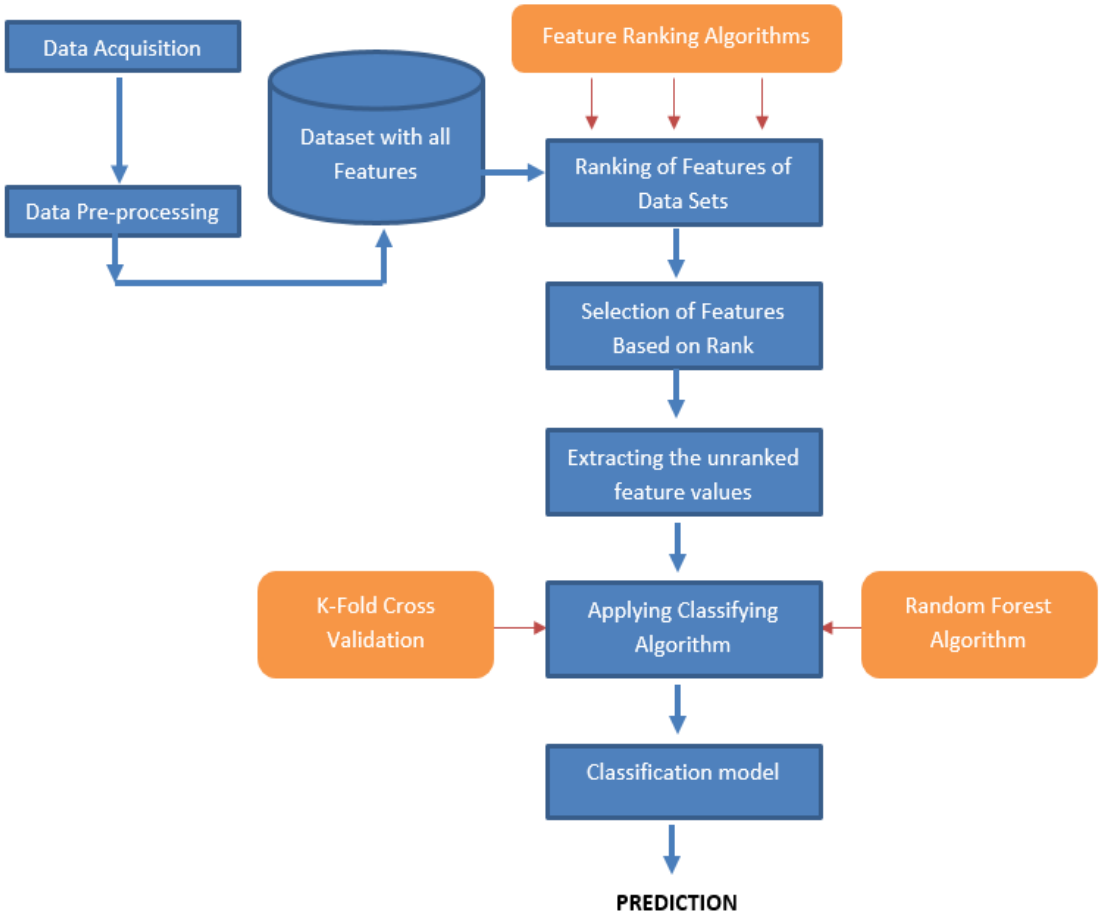

Fig 7. Architecture of the Prediction Model using Ranking Algorithms and Classifier Algorithm

5.1 Data Acquisition and Pre-processing:

A large data set is required for making a clear and accurate prediction. The National Center for Biotechnology Gene Expression Omnibus database was referred. The database included molecular cell details for peripheral blood monocular class and bone marrow studies as shown in Fig 8.

The datasets were divided into three groups namely-HG-U133A microarray(dataset1), the HG-U133 2.0 microarray(dataset2), and Illumina RNA-seq (dataset3) (Warnat-Herresthal, et al., 2020) base don three platforms as shown in Fig 7. The samples included in the dataset were mainly Acute Lymphocytic Leukaemia(ALL), Acute Myeloid Leukaemia(AML),Chronic Lymphocytic Leukaemia(CLL), Chronic Myeloid Leukaemia(CML) , Myelodysplastic syndrome(MDS) and other non-leukaemia diseases as shown in Fig 9 [19]. Duplicate values of the data sets were excluded and pre-filtered.

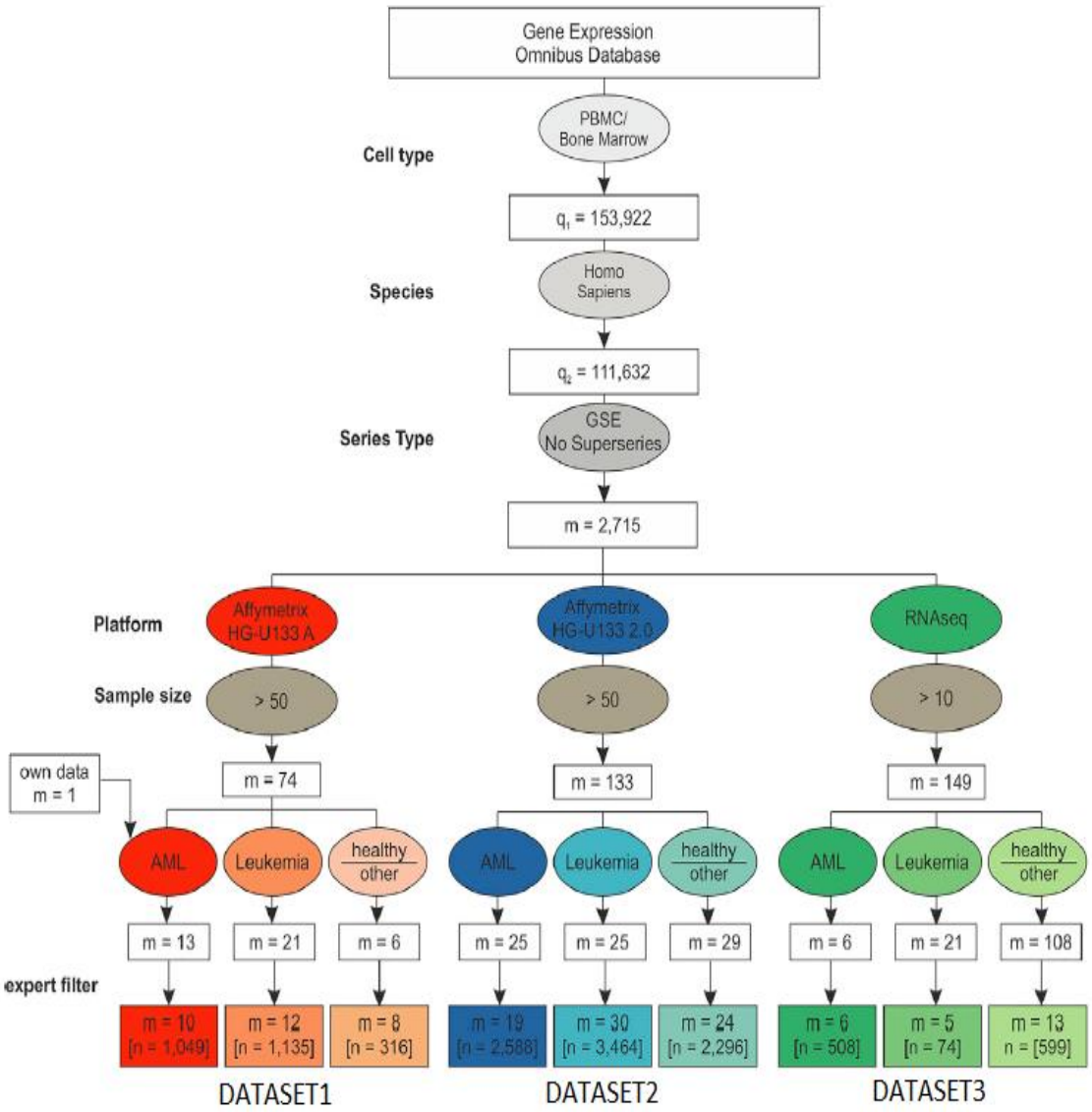

Fig 8. Dataset for classification of blood cancer stages [19]

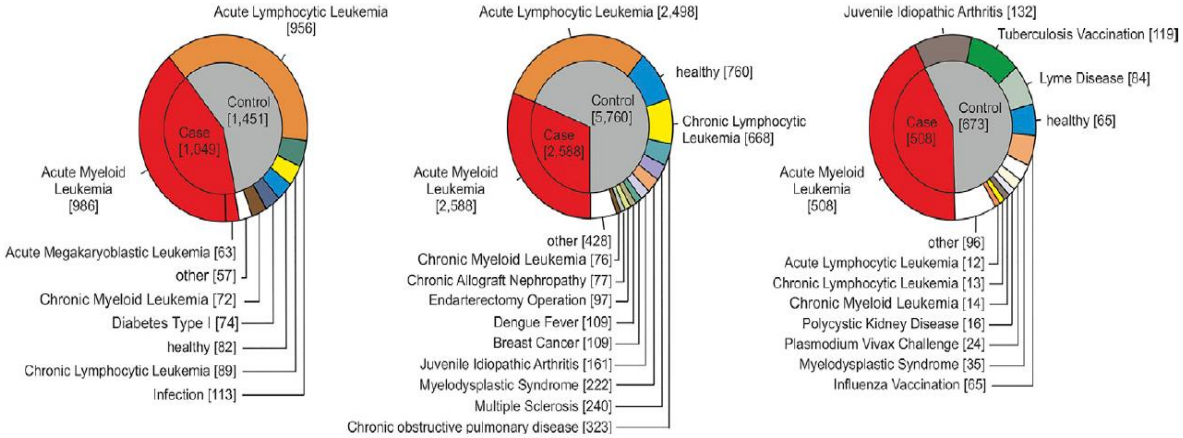

Fig 9. Dataset containing various Cancer Types [19]

The datasets created will also have information regarding the various stages of blood cancer. The information are regarding the cell details. This cell details are very much helpful in classification of stages in blood cancer.

5.2 Ranking of features and selection of features

Following the data collection and pre-processing, the data have to be now ranked and then selected. The ranking is based on certain features of the data sets. Ranking algorithms are used for this step. An amalgamation of several search techniques are utilized for feature selection and the result is a feature subset. Additionally a scorecard is maintained to keep track of evaluation process.

Machine learning literature proposes a variety of feature ranking and selection methods. During this stage, irrelevant features are discarded. In the architecture, we propose two steps for the feature ranking and selection. First step is the subset generation followed by a subset evaluation stage as depicted in Fig 10 [20]. Subset evaluation stage involves a filter method. The Table 1 below gives certain performance domains based on which the evaluation and ranking of feature subsets are done [20].

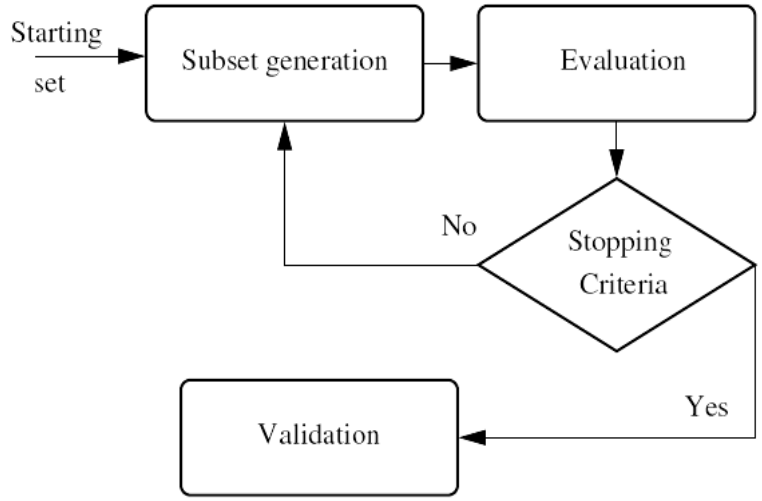

Fig 10. Feature Selection-Subset creation and Subset evaluation [20]

Table 1. Evaluation Attributes

S.No	Attributes
1	Information Gain Attribute Evaluation-IG
2	Gain Ratio Attribute Evaluation-GR
3	Symmetrical Uncertainty Attribute Evaluation-SU
4	Relief-F Attribute Evaluation-RF
5	One-R Attribute Evaluation-OR
6	Chi-Squared Attribute Evaluation-CS

The ranking is done on additional factors such as methods used frequently and entropy-based and statistical data.

The entropy is the measure of purity of samples and is very important as it is embedded in the basics of IG, GR and SU attributes. For a variable Y , the entropy is given by

$$H(Y) = -\sum_{y \in Y} p(y)\log_2(p(y))$$

Where *p(y)* is the probability density function for the variable Y. The entropy of Y after X is added to it.

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 (p(y|x))$$

Where *p(y|x)* is the conditional probability of *y* given variable *x.*

Information Gain (IG) is the measurement of additional information about Y given by X which is an indication Y level entropy decreases.

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

In order to compensate for the bias of IG a non-symmetrical measure called Gain Ratio (GR) is introduced.

$$GR = \frac{IG}{H(X)}$$

Symmetrical Uncertainty (SU) is given by

$$SU = 2\frac{IG}{H(Y) + H(X)}$$

For calculating, the worth of a particular feature chi-squared method is the most common method. It is given as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Oij - Eij)^2}{Eij}$$

Where observed frequency is O_{ij} and the anticipated frequency is E_{ij}.

When using OneR, for each attribute a single rule is created. After this from all the rules, the attribute with the smallest error is selected. Another method termed as Relief-F uses repeated sampling method of sampling a particular instance several times and then taking into account features with values which are nearest to each other.

5.3 Classification Stage and Model Creation

The input for the classifier algorithm for detecting and classifying the various stages of blood cancer are three datasets as explained in the section 5.1 and shown in Fig 6. Therefore, three models for each dataset have to be constructed. Additionally a base model that ignores all the feature ranking and feature selection is also constructed. All the features of the entire dataset will be considered for the construction of the base model.

For the Purpose of classification, a number of machine learning algorithms are available. Among them, the most common ones are IB1, Naïve Bayes, Random Forest and Radial Basis Function. IB1 employs the concept of Euclidean distance and is considered as the nearest neighbor classifier. The idea behind IB1 classifier is that it attempts to find the training instance nearest to the test data and then the prediction is made relative to the training data. Bayes' theorem is the basis for the Naïve Bayes algorithm. A higher and efficient performance is achieved through the use of Naïve Bayes. In Naïve Bayes algorithm, assumptions are made considering that all the features not nondependent and this

makes learning more simplified. Radial Basis Function is a feed forward network consisting of input layer and two layers.

One of the most used classifier compared to the counterparts is the random forest that is supervised learning algorithm. Random forest algorithm is an amalgamation of several decision trees as shown in Fig 11. The random forest decision trees are trained using bagging method. The principle behind this is that the overall result could be alleviated by the consolidation of several learning models. Compared to the other classifiers, following are the advantages of using random forest as the classifier in the proposed model for detection and classification of various stages of blood cancer:

- Random forest has the feature coupling classification and regression.
- Predictions can be done on variable data.
- Works very well with data that are missing.
- Higher accuracy is provided as several decision trees are involved.
- Able to knob big data with diverse variables.

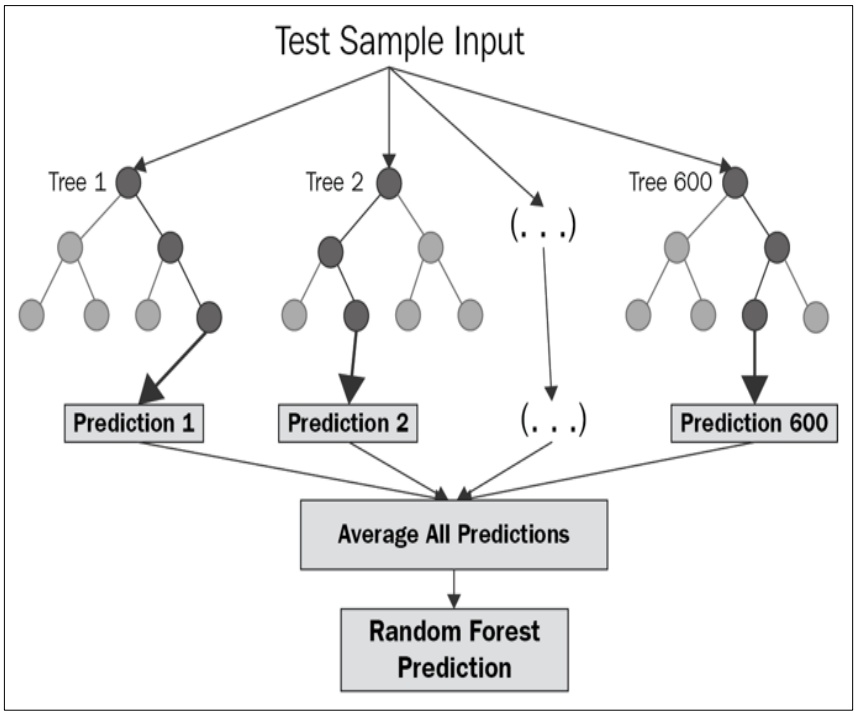

Fig.11 Random Forest with two Trees [13]

When compared to normal decision trees, the random forest does not suffer from overfitting. Moreover, in comparison to decision trees, the random forest randomly makes selection of features and builds random forest and average results are considered.

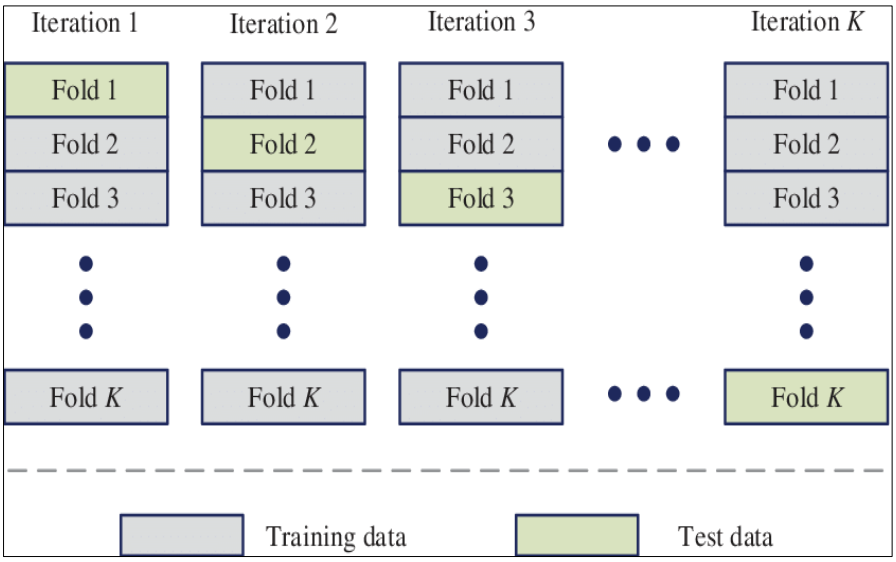

Fig 12. K-Fold Cross Validation Using randon forest

The proposed model makes classification and trained with the datasets a model is created.A K-Fold cross validation process is applied and used for testing as shown in Fig 12. By using random forest along with K-fold cross validation a reliable, accurate and stable predictions are achieved.

6 Results and Discussions

In order to perform the task a windows 10 operating system with 16GB RAM with an Intel Core i7 processor was set up. Ranking methods IG,CS,RF,GR, OR and SU were employed for performing ranking of the selected features.

Table 2. Result of Ranking Methods on Features selected in Dataset

Features-F	IG	CS	RF	GR	OR	SU
F1	10	10	9	10	10	9
F2	11	11	11	10	10	11
F3	7	9	10	6	11	10
F4	8	8	8	9	10	9
F5	1	1	1	1	3	2

As discussed in section 5.1 four models were created-one base model and three models based on datasets.For all the mdels created a K-Fold Cross Validation is performed. Base don the ranks in each model one feature is eliminated.After this, the best performing model is selected out of the four created models.The next step is to perform the independent test on the best performing model selected and base model.

Several features from the data sets were selected and then a comparison was made with the blood samples of cancer patients. Using this data, a model was created and a classification based on the test datasets in reference was made about the stages of blood cancer. Summarising this data, the classifier classified the AML blood samples and patients' blood cancer state is predicted. Moreover, advanced

techniques could be embedded in the model and can be used for predicting not only the blood cancer stages but also the survival rates and reoccurrence of the blood cancer in patients.

7 Challenges and Future Works

The dataset used as the training data and test data were extracted from The National Center for Biotechnology Gene Expression Omnibus database. This database was created for multipurpose and possess various designs and aims. In addition, in order to better the model a more significant understanding of the machine learning techniques and haematology is essential. As medical diagnosis and treatment go hand in hand, a thorough study of side-effects is also needed. This is due to the fact that the classifier gives variant results depending on the input provided. Though the random forest classifier gives an average classification based on decisions of several decision trees, the performance depends on the test data and training data. Another aspect to be considered here is that a continuous monitoring of the model is required and the performance has to be kept always in check even after the model development.

Despite the fact that artificial intelligence and the techniques involved have created a huge difference in the prediction of cancer diagnosis and treatment, quite a few bottlenecks and stumbling blocks extant. As a result of numerous research, plentiful information and data are provided by health care suppliers due to the improving diagnosis tools. Quality, access and recovery of these data needs to be guaranteed by benchmarking it and standardising the use. As part of quality assurance and labelling and annotations the datasets stored in databases are frequently updated. The data curation process happens very often with new innovations and this is a major hindrance in the usage of consistent data as it requires trained experts and professionals. This may result in lot of time and cost may also increase.

8 Conclusion

Machine learning has proved to be a game changer in the field of blood cancer prediction, diagosis and treatment prescribing. However, as the training data and test data varies the predictions may also change at various instances. Radiology at certain situations may fail to provide significant results because of imaging techniques employed. Medical field has to go beyond expectations and deliver the diagnosis and treatment for patients at an earlier stage.Mecial imaging such as scanning has to inprove the quality of images provided so that an accurate signs could be given regarding athe abnormality of the cells on the body.Genetic predisposition to cancer can be predicted using heriditary data.Applying artificial intelligemce could certainly pave way for advanced genomiccs and cancer screening.Issues that concern the use of artificial intelligence are numnerous but not limited to securoty concerns, data availability , specialisations etc.

References

1. Al-Shamasneh, A. R. M., & Obaidellah, U. H. B. (2017). Artificial intelligence techniques for cancer detection and classification: review study. *European Scientific Journal*, *13*(3), 342-370.
2. Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., & Mak, R. H. (2019). *Artificial Intelligence In Cancer Imaging: Clinical Challenges And Applications*. CA: a cancer journal for clinicians, 69(2), 127-157.
3. Chaurasia, V., & Pal, S. (2017). *A Novel Approach For Breast Cancer Detection Using Data Mining Techniques*. International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol, 2.
4. Eshtay.M, Faris.H, & Obeid.N. (2018). Improving exteme learning machine by competitive swarm optimisation and its application for medical diagnosis problems. *Expert System Application*.
5. Gupta, S., & Kumar, D. (2018). Pattern Classification of Breast Cancer Patients for Personalized Medical Diagnosis. *IJLTET–International Journal of Latest Trends in Engineering and Technology*, *11*.

6. Hirasawa, T., Aoyama, K., Tanimoto, T., Ishihara, S., Shichijo, S., Ozawa, T., ... & Tada, T. (2018). Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. *Gastric Cancer*, *21*(4), 653-660.
7. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.
8. Liu, C., Pan, C., Shen, J., Wang, H., & Yong, L. (2011). MALDI-TOF MS Combined With Magnetic Beads For Detecting Serum Protein Biomarkers And Establishment Of Boosting Decision Tree Model For Diagnosis Of Colorectal Cancer. International journal of medical sciences, 8(1), 39.
9. Lopez-Rincon, A., Tonda, A., Elati, M., Schwander, O., Piwowarski, B., & Gallinari, P. (2018). *Evolutionary Optimization Of Convolutional Neural Networks For Cancer Mirna Biomarkers Classification*. Applied Soft Computing, *65*, 91-100.
10. Manogaran, G., Vijayakumar, V., Varatharajan, R., Kumar, P. M., Sundarasekar, R., & Hsu, C. H. (2018). *Machine Learning Based Big Data Processing Framework For Cancer Diagnosis Using Hidden Markov Model And GM Clustering*. Wireless personal communications, 102(3), 2099-2116.
11. Mocan, I., Itu, R., Ciurte, A., Danescu, R., & Buiga, R. (2018). *Automatic Detection of Tumour Cells In Microscopic Images Of Unstained Blood Using Convolutional Neural Networks*. In 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP) (pp. 319-324). IEEE.
12. Ndivya. (2018, November 16). AI-based diagnostic test to identify cancer at earliest stages. Retrieved from https://www.medicaldevice-network.com/news/ai-based-diagnostic-test/.
13. Niklas Donges(2020), A Complete Guide to the Random Forest Algorithm, Retrieved from https://builtin.com/data-science/random-forest-algorithm
14. Payandeh, M., Aeinfar, M., Aeinfar, V., & Hayati, M. (2009). A new method for diagnosis and predicting blood disorder and cancer using artificial intelligence (artificial neural networks). *International Journal of Hematology-Oncology and Stem Cell Research*, *3*(4), 25-33.
15. Sayed, S. (2018, December 10). Machine Learning Is The Future Of Cancer Prediction. Retrieved from https://towardsdatascience.com/machine-learning-is-the-future-of-cancer-prediction-e4d28e7e6dfa.
16. Subbaiah.S,Muruganandam.S(2020),Applications of Machine Learning in Cancer Prediction,*International Journal of Engineering Research and Technology,Vol 8,Issue 03.*
17. Towers-Clark, C. (2019, May 14). The Cutting-Edge Of AI Cancer Detection. Retrieved from https://www.forbes.com/sites/charlestowersclark/2019/04/30/the-cutting-edge-of-ai-cancer-detection/#49844ebe7336.
18. Vogado, L. H. S., Veras, R. D. M. S., Andrade, A. R., De Araujo, F. H. D., e Silva, R. R. V., & Aires, K. R. T. (2017). *Diagnosing Leukaemia in Blood Smear Images Using an Ensemble of Classifiers and Pre-Trained Convolutional Neural Networks*. In 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) (pp. 367-373). IEEE.
19. Warnat-Herresthal, S., Perrakis, K., Taschler, B., Becker, M., Baßler, K., Beyer, M., L.Schultze, J. (2020). Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics. *iScience*.
20. Novakovic,J.,Strbac,P.,Bulatovic,D.(2011),Toward Optimal Feature Selection using Ranking Methods and Classification Algorithms,*Yugoslav Jpournal of Oeprations Research,Vol 21,Issue* 1,pp. 119-135