National College of Ireland

Prediction of Malignant Melanoma using Machine Learning

MSc Research Project
MSc Data Analytics

Elizabath Kaimoolyil Thomas
Student ID: X20170131

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Elizabath Kaimoolayil Thomas		
Student ID:	X20170131		
Programme:	MSc. Data Analytics	**Year:**	2021.
Module:	Research Project		
Supervisor:	Jorge Basilio		
Submission Due Date:	31/01/2022		
Project Title:	Prediction of Malignant Melanoma using Machine Learning		
Word Count:	6571	**Page Count:**	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Elizabath Kaimoolayil Thomas……

Date: 16/12/2021……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Malignant Melanoma using Machine Learning

Elizabath Kaimoolayil Thomas
X20170131

Abstract

Tumour is one of the most rapidly spreading and severe diseases that many people are dealing with nowadays. Among this melanoma is most scary skin cancer that affected by many of the people in these days. So early prediction of this malign and benign melanomas better for the early recovery of the patients. Toady machine learning application in health care is highly increasing. So, for the prediction of malign melanoma in this paper we used 3 different machine earning methods. And analysed each model with different evaluation methods. In this study SVM, Logistic regression and random forest are used as the prediction models. Here the random forest predicts the target with that having accuracy of 74% which has the highest accuracy among the 3 models. SVM and Logistic Regression predict the outcome in accuracy of 63 and 65 respectively. The precision, recall, f1score and area under the ROC curve also shows that random forest is the best among these and other two have an average performance.

Key Words: machine learning, melanoma, prediction, colour

1 Introduction

The quality of healthcare data influences every decision made during the patient care process. The value of reliable and accurate data has increased dramatically. There has been a significant increase in the analysis of healthcare data in recent years. Healthcare practitioners are increasingly using sophisticated communication and networking systems throughout their settings due to technological advancements and other factors. Analytics is a tool or set of approaches for converting raw data into meaningful and useable information, which is then used to assist healthcare organizations in making effective strategic and operational choices and defining their long-term vision. Healthcare organizations are increasingly turning to analytics to acquire fresh insights from their data. New analytics techniques are being used to develop clinical and operational breakthroughs in order to address business difficulties. Predictive analytics in healthcare involves revealing strategies to assist organizations foresee better future possibilities, construct better health services, guard against fraud, and forecast patient status.

Many people are now afflicted with various types of cancer. Melanoma is a kind of cancer find in skin that occurs when melanocytes create pigment. Melanomas are malignant tumors that form on the surface of the skin. Those with low amounts of melanin, a skin pigment, are at unable to produce melanomas if they are exposed to UV radiation. Melanoma is very dangerous type of skin cancer. In 2012, a total of 232,000 cases occurred over the world. In 2015, the Centers for Illness Control and Prevention (CDC) reported 3,1 million

active illness cases. Melanoma is more common in Australia and New Zealand than everywhere else in the world. In United States, males are more affected this likely than females, with a ratio of roughly 1:6. As a result, it is crucial to discover this malignancy and its severity early on. Machine learning is being used into today's technology breakthroughs in the health profession. This facilitates in the assessment of a wide range of health issues.

In this study we are trying to predict the malign melanoma and benign melanoma by using machine learning techniques. We are trying to make most accurate machine larning model for the prediction. So after analysing different machine learning algorithms, in this research we are using support vector machine, random forest classifier, and logistic regression as our prediction algorithms. All these are classification models because our data is binary classification data frame. Apart from other studies here we are using 2 different type of data one which contain images of the cancerous part and one csv file that contain different attributes like sex, age, diagnosis, patient id, image name etc. so we are trying to find and obtain various features from the images and append it to the data frame. By this way we can analyse how other categories related to benign and malign melanoma. These 2 datasets are available in Kaggle. We have referred some previous works related to this study and came with this research question that try to find the malignity using both type of datasets. From this research people can understand how to handle both image dataset and tabular data in Data analysis.

2 Related Work

The main purpose of this work [1] by Waheed et al. (2017) is to develop an effective machine-learning approach for identifying melanomas using dermoscopic images. It categorizes melanomic sores according to their distinct features. Depending on recognizable traits and varying levels of melanomic lesions, this approach begins by recognizing color - shape cues from dermoscopic pictures. Melanoma dermoscopic images are categorized when they are provided to a classification model. A section on melanomas, as well as colour and textural features, is included. The suggested approach is tested on the PH2 dataset.

In this research [2], Vijayalakshmi (2019) proposes a completely automated solution to dermatological disease identification utilizing lesion photos, as opposed to earlier medical equipment detection. This model is broken down into three stages: data collection and enhancement, model creation, and prediction, to name a few. To create a better structure, image processing methods were coupled with AI algorithms, resulting in an 85 percent increase in accuracy. This study used scan pictures to predict melanoma tumours.

In this research article [3] Using a dataset of 1227 UM patients, Cox proportional hazard analysis was used to investigate important predictors of metastasis, including medical and chromosomal characteristics. The researchers employed machine learning techniques such as logistic regression, decision trees, survival random forest, and survival-based regression methods to evaluate a multidimensional strategy for predicting metastatic risk. According on cross-validation results, a logistic regression classifier was created to calculate a tailored risk of metastasis based purely on clinical and chromosomal factors. The risk prediction accuracy ranged from 80% (using just chromosomal data), 83 percent (using only clinical variables

such as age, sex, tumor location, and size), and 85 percent (using both chromosomal and clinical features) (clinical and chromosomal information).

In this work, Mocellin et al. (2006) used support vector machines (SVMs) to explore the predictive potential of data analysis [4]. The performance of the SVM-based SLN status estimation was evaluated using logistic regression. As per recent study, SVM-based sentinel lymph node (SLN) status forecasting could be utilized as a forecast strategy to prevent sentinel lymph node biopsy (SNB) in 60percent of individuals who are currently eligible, with a good precision. If this treatment could be verified in a wider group of patients, it would benefit both their quality of life and the medical system's expenses. This study employs a machine learning technique to predict the SLN state. As a consequence, understanding how machine learning techniques are employed in diagnosis much simpler.

In this paper [5] Grzymala-Busse et al. (2001) employed the LERS data-mining method to validate over 20,000 revised ABCD equations, and the estimated error rate of melanoma diagnosis was determined for each updated formula using 10-fold cross-validation. As a result, they came up with the optimal ABCD formula for our setup: cluster analysis-based discretization, the LEM2 algorithm, and the traditional LERS classification scheme. The ABCD formula is advantageous, according to the research, because the percentage error without it was greater for the data set (13.73 percent). This study only looks at the melanoma diagnostic mistake rate. As a result, it aids in the detection of problems, allowing us to correct them.

The purpose of this research[6] is to find serological indicators that may be used in relation to clinical and histological characteristics of the disease to forecast spreading incidence in people who are still in the early stages of infection. Mancuso et al. (2020) devised a framework able of correctly diagnosing initial melanomas with a high or low risk of spreading using ELISA and Luminex evaluations of chosen cytokines, as well as machine learning and Kaplan–Meier algorithms for data processing.

In this work [7] Mccarthy et al. (2004) conducted a comprehensive review of the many types of machine learning techniques that are being used, the sorts of data that are being included, and the quality of these approaches in cancer detection and diagnosis. This article aids in the gathering of general information on the identification of various malignancies. so that we can distinguish melanoma tumors from other cancers, which will aid in the clarity of our research. Furthermore, the developed algorithm serves as a novel tool for distinguishing patients with a high success rate from those who are at high risk of future metastasis.

There are presently no clinically appropriate indicators for immune checkpoint inhibitors, according to Johannet et al. (2021). (ICIs). They employed histological data and health information in this work [8], and deep learning has been used to estimate ICI reaction in metastatic disease. based on neural network forecasts and medical studies, created a multi-variable predictor To designate clients as high- or low-risk for advancement, a ROC curve was created and the appropriate threshold was chosen. To evaluate PFS between the categories, Kaplan–Meier curves were used. 2nd generation scanners, the Aperio AT2 and the Leica SCN400, were used to test the predictor. The multivariable classifier forecasted response with AUC 0.800 and AUC 0.805 on images from the Aperio AT2 and images from the Leica SCN400, respectively. Patients were correctly classified as having a high or low

risk of illness development using the classifier. PFS was substantially lower for Those patients categorized as high risk for advancement than for those classified as low risk.

This paper [9] by Tsur et al. build a new customization method that might be helpful for the clinical context for forecasting the duration to disease improvement in patients taking pembrolizumab. The connections of a progressed melanoma tumor with the innate immunity and the immunotherapy medication pembrolizumab were studied using a basic mathematical model. They converted the framework into an algorithm which can forecast a person's specific reaction to a drug when paired with medical background information. The accuracy of the system's predictions was tested by using Leave-One-Out cross-validation method. Among the clinical parameters evaluated, early tumor burden, Breslow tumor size, and the existence of multilayer melanoma also were found to be highly associated to the activating rate of CD8+ T cells as well as the gross tumor pattern of increase. They used the measurements of these correlates to tailor the statistical model to estimate the length to progression of individual patients (Cohen's = 0.489). A comparison of the expected and actual time to advancement in patients who progressed during the follow-up period demonstrated moderate accuracy (R2 = 0.505).

This paper [10] describes how machine learning is often used in health sciences to date and provides an extensive literature review on the subject. They did a comprehensive analysis of the literature, locating the material using a PubMed database search. Assessment and qualification were evaluated by two clinicians based on pre-determined evaluation criteria. The study found that in dermatology, machine learning technologies were tried in eight main classifications. The majority of the systems used image recognition software geared mainly at binary classification of malignant melanoma (MM). Tables offer brief system descriptions and results. In each of the eight categories, impressive results were recorded, but a head-to-head comparison proved hard. The fact that we found machine learning techniques in so many different areas of dermatology demonstrates how diverse machine learning is.

Kniep et al. (2019) presented quantitative aspects of conventional brain MR images that were employed in a machine learning classifier to detect the tumor level of brain tumors, with excellent discriminating accuracy. A prototype fivefold cross validation approach was used to validate the results. The consistency of the findings was tested by comparing 10 cross-validation sets chosen at random. The classification findings were compared to predictions made by two radiologists based on a traditional reading. These values ranged from 0.64 for non–small cell lung cancer to 0.82 for melanoma in the five-class problem, with all P values below.01. The classifier's accuracy was higher than the radiologists' findings. Melanomas showed a 17-percentage-point increase in sensitivity when compared to both readers' sensitivity; Significance level were less than.02.

In this paper [12], Mccarthy et al. (2004) explore how to obtain therapeutically significant information from a diverse collection of genetic data using exploratory data analysis approaches such as machine learning and high-dimensional visualization. The paper then goes on to describe two sophisticated algorithms (PURS and RadVizTM) that they have found to be useful in exploratory data analysis of large biological data sets. We next use three distinct forms of molecular data to show how these tools may be used to identify, diagnose, and manage cancer. They begin by discussing how our exploratory analytic approaches may be used to detect ovarian cancer using proteome mass spectroscopy data. They then go into

how these approaches may be used to distinguish between squamous and adenocarcinoma of the lung using gene expression data for diagnostic purposes. Finally, they show how to utilize such algorithms to choose the most effective chemical compounds from a collection of thousands to treat patients with melanoma versus leukemia.

In this paper, Mccarthy et al. (2004) [13] show how exploratory research approaches like machine learning and high-dimensional visualization may be used to retrieve therapeutically valuable information from a variety of genetic data. They explain two proprietary algorithms (PURS and RadVizTM) that they have found to be beneficial in the experimental exploration of huge biological data sets after an introduction to machine learning and visualization techniques. They then use three distinct forms of molecular data to demonstrate the relevance of these methodologies to cancer detection, diagnosis, and management using three separate instances. They begin by discussing the identification of ovarian cancer using our exploratory analytic techniques on proteomic mass spectroscopy data. They then go into how these approaches may be used to discern between squamous and adenocarcinoma of the lung using genetic data. Finally, demonstrate how to apply such algorithms to choose the most effective chemical compounds from a database of chemical compounds for individuals with melanoma vs leukemia.

Harbour (2014) developed a gene expression profile (GEP) that distinguishes between early uveal melanomas with a low metastatic risk (class 1 tumors) and those with a high metastatic risk (class 2 tumors) in this study [14]. (class 2 tumors). They collaborated with a number of centers to show that their specimen collection system was straightforward to learn and use, and that it allowed samples to be transported safely and reliably from remote locations with a low failure rate. Finally, they showed that our GEP assay surpassed the previous gold standard, chromosome 3 monosomy 3 testing, in multicenter prospective study in predicting which individuals will develop metastatic sickness. This is the only prognostic test for uveal melanoma that has ever undergone such thorough validation, and it is now being utilized in over 100 locations in the United States and Canada under the brand name DecisionDx-UM.

In this current work [15] Kawahara, et al. used a machine learning (ML) technique with radiomics characteristics to develop a method to forecast the reaction of brain metastases (BMs) cured with Gamma knife radiosurgery (GKRS). The local response (LR) of 157 metastatic brain tumors was divided into two groups using MR imaging data obtained using a FLASH (3D rapid, low-angle shot) scanning technique with gadolinium (Gd) contrast-enhanced T1-weighting. They ran a radiomics study on the tumors, which yielded over 700 characteristics. A neural network (NN) classifier with 10 hidden layers and rectified linear unit activation was used to create a prediction model. Five-fold cross-validation was used to assess the training model. The NN model was applied to a set of data that was not utilized to create the model for the final assessment. The accuracy and specificity of the LR forecasting model, as well as the area under the receiver operating characteristic curve (AUC), were investigated. The visual evaluation method's accuracy and sensitivity were 44 and 54 percent, respectively. The suggested NN model, on the other hand, has accuracy and sensitivity of 78 and 87 percent, respectively, and an AUC of 0.87.

According to Abraham et al. (2021) [16] , gene expression study was utilized to detect the origin area alone, but it suffers from low neoplastic percentage in stages of the disease, where identification is often necessary. MI was utilized by them. GPSai is a genomic Prevalence

Score that uses machine learning to integrate DNA sequencing and complete gene information to aid in cancer identification. The system was trained on 34,352 cases of genomic data and 23,137 cases of genomic and transcriptomic data, and it was verified on 19,555 cases. While deciding between 21 different cancer categories, MI GPSai identified the tumor type in the labeled data set with an accuracy of above 94 percent in 93 percent of cases. When the second highest forecast is taken into account, the accuracy rises to 97 percent. In addition, MI GPSai was able to predict 71.7 percent of CUP instances.

Shoombuatong et al. employed an interpretable random forest classifier, as well as amino acid composition, dipeptide structure, and pseudo amino acid composition, to develop THPep, a sequence-based approach for predicting and evaluating tumor homing peptides. On an objective benchmark dataset, a second test set achieved an overall accuracy of 90.13 percent and a Matthews correlation value of 0.76. According to the findings, THPep beat the current approach and has a lot of potential as a model for calculating tumor attractive peptides.

Jen Tseng et al. evaluated three machine learning algorithms in this research [18] to find crucial risk indicators to determine the recurrence-proneness of cervical cancer: support vector machine, C5.0, and extreme learning machine. The Chung Shan Medical University Hospital Tumor Registry was given access to the medical data and pathology. The C5.0 model is the most effective method for identifying recurrence-proneness variables, according to the findings of experiments. According to the data, Pathologic Stage, Pathologic T, Cell Type, and RT Target Summary were the four most important recurrence-proneness characteristics. Pathologic Stage and Pathologic T, in particular, were significant and independent predictors of prognosis. Clinical trials should randomize patients stratified by these prognostic characteristics to examine the effectiveness of adjuvant therapy. Additionally, better surveillance following treatment may lead to earlier discovery of recurrence and more exact evaluation of recurrent status, which may improve prognosis.

All of the publications listed above have aided me in determining my study question and topic. Melanoma is a dangerous illness, as evidenced by research, and early detection is critical for medical professionals to provide better treatment. The use of machine learning techniques in the health industry is clearly rising, as seen by the publications mentioned above.

3 Research Methodology

It is clear from the preceding literature analysis that several Machine Learning approaches are accessible. As a result, choosing the best model is critical. This is accomplished via the use of several methods that help in the extraction of data-related information. Data collection, data cleaning, data processing, data exploration, and data mining are all phases in this process. It's also crucial to pick the ideal tool for detecting distinct trends in the data. This study makes use of the Knowledge Discovery Database (KDD). It begins with the raw data in the data source. The process that will take place in the research is plotted in the below flowchart figure1.

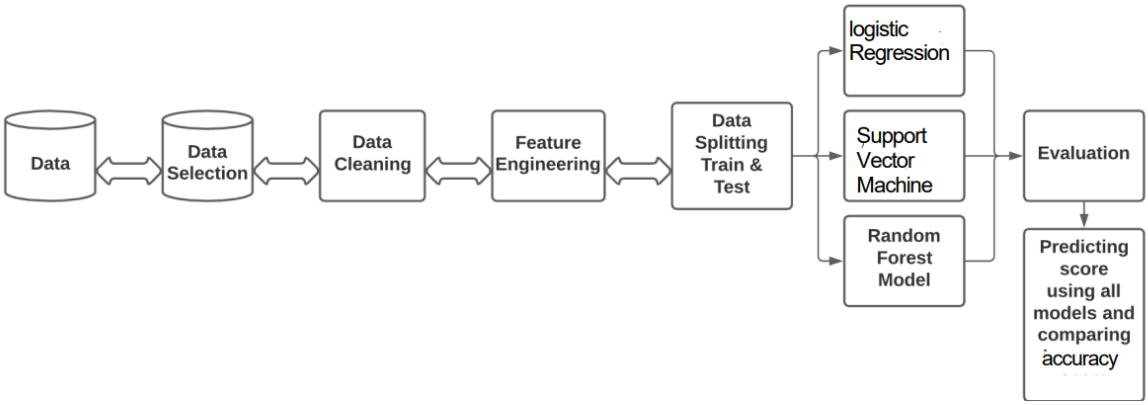

Figure 1. Research Framework

3.1 Data Gathering

Data collection is critical in MI to make successful outcome. The raw data in our study is the data files that we got from Kaggle. Which is a great open source of datasets for data science research. We are using 2 types of datasets. One which is an image dataset contain images of different melanoma tumours. And the second one is a CSV file which contain various attributes related to melanoma. The dataset comprises 33,126 dermoscopy photographs or training images of different malignant or benign skin conditions from over 2,000 patients. Each image is connected to these other persons using a specific patient identification. All malignant assessments were confirmed with histology, while benign diagnoses were confirmed with facilitates, longitudinal follow-up, and histopathology. The International Skin Imaging Collaboration (ISIC) created the collection, which includes photos from the Hospital Clnic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, University of Queensland, and University of Athens Medical School. Images are also provided in JPEG format and csv file as well. The description of each columns are given below.

- image_name - separate identification ids of images,
- patient_id - patient identification number
- sex – patient's sex
- age_approx – average age of patient
- anatom_site_general_challenge – image location in the body
- diagnosis – diagnosis of each image
- benign_malignant – malign or benign images
- target - The target variable that was already converted into a binary.

3.2 Data pre-processing and Cleaning

Data pre-processing and data cleaning is the first step in the KDD. In our dataset we have a huge umber of null values. I used dropna() function and fillna using the average value to

remove the null values. Then removed the unwanted columns like patient id, diagnosis, image name from the data. Because the all factors don't have any importance in the model building. Also convert the string values in the dataset to numerical values. This step required for the model building. In the variable diagnosis most of the features are unknown. This will affect the accuracy of our model. So removed the diagnosis feature from the data for the better prediction.

In the image data we resized the image to (256, 256) we have a very large data and it is better to use small size for the analysis. The next step is to extract features from images and insert them to the dataframe. In our research, we attempted to extract the tumor's picture color and size. However, color detection was successful. In the case of size, only a picture with a reference object can be used to determine the object's size. Using the webcolors package, we extract the average color of each image and append it to the dataframe containing melanoma patient characteristics. The given picture below shows the final form of dataframe.

Our dataset is highly unbalanced so we use over sampling to make it balanced. And then we got the final data frame as shown below in the figure 2.

```
melanoma_df
```

	sex	age_approx	anatom_site_general_challenge	target	color
6454	2	45.0	3	1	1.0
25238	2	25.0	2	1	1.0
16139	1	55.0	3	1	1.0
6628	2	35.0	1	1	1.0
4449	1	65.0	4	1	0.0
...	...	...	...	...	...
33121	1	50.0	1	0	1.0
33122	1	65.0	1	0	0.0
33123	1	20.0	2	0	1.0
33124	1	50.0	2	0	1.0
33125	1	45.0	1	0	1.0

63912 rows × 5 columns

Figure 2. final form Datafrmae

3.3 Visualization.

Visualizing the data frame with bar graphs, pie charts, histograms, and other tools is highly useful for acquiring information, trends, and understanding about the dataset. It is one of the most well-known data analysis processes. This is really useful for comprehending and explaining facts to others.

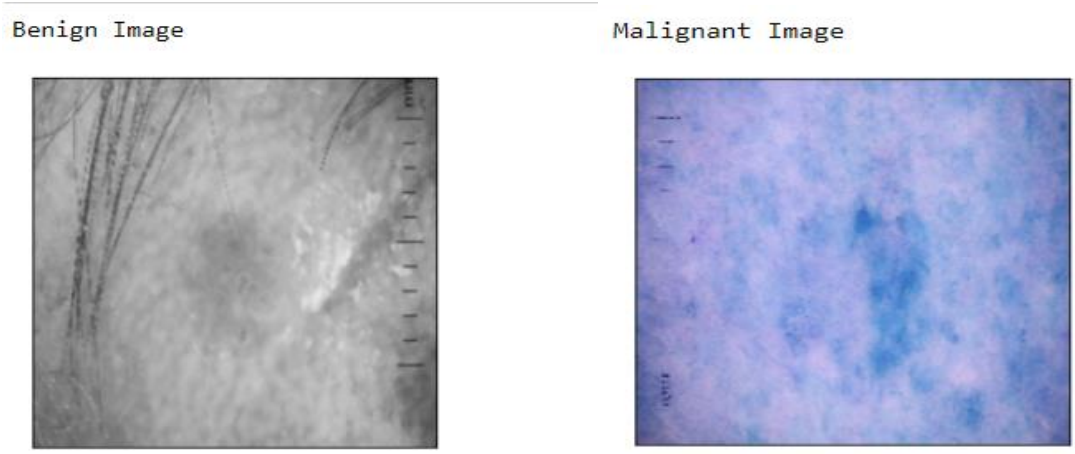

Figure3. Benign and Malignant melanoma

In the image dataset we have malignant and benign melanoma images. The above figure2 shows the images of benign and malignant melanoma.

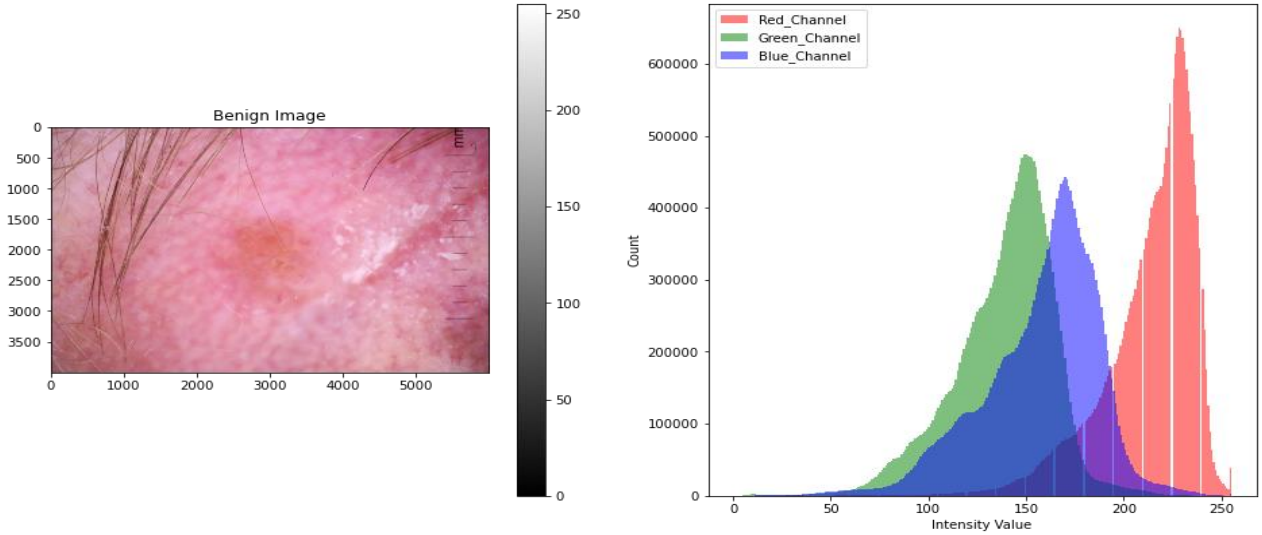

Figure 4. Intensity of RGB vales in an Benign image

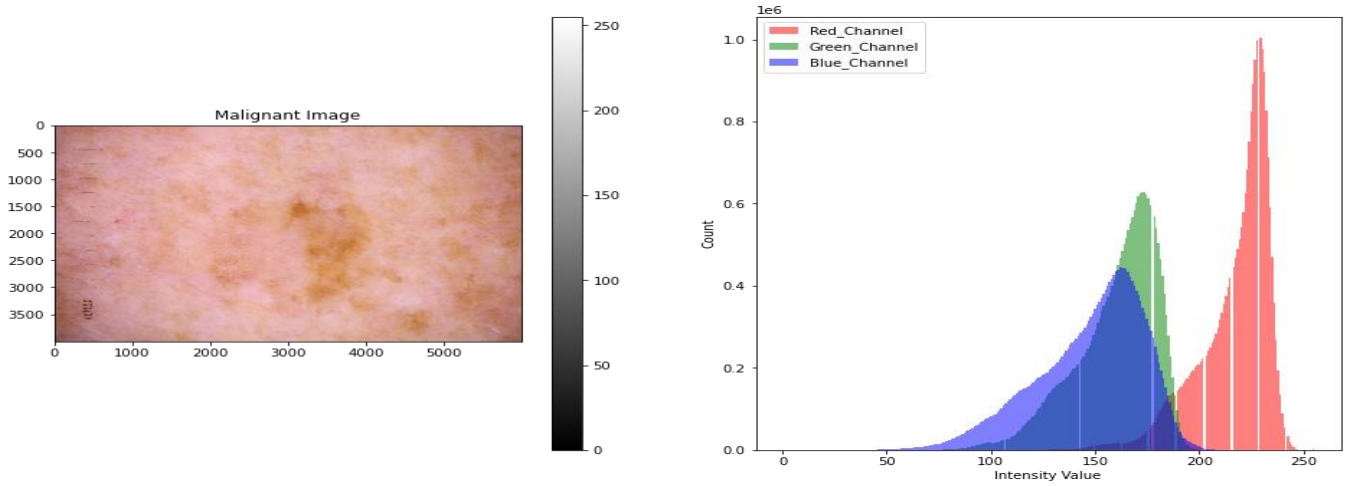

Figure 5. intensity of RGB values in malignant image.

The above 2 figures (figure4 and figure 5) shows the intensity of RGB (Red, Green, Blue) values in Benign image and malignant image. It is clear from both the image that red channel has the high intensity. But the values are slightly different. The 2 images are randomly selected and the values will change according to the images.

The given below 2 pie chart shows ratio of anatom_site_general_challenge that is the location of the images and the diagnosis. That means the condition and area affected by the condition. From the figure 6 it is clear that in our dataset most of the images are of torso. And only a very few images of oral. And from figure 7 it is evident that most of the images are of the condition nevus and the least is the atypical melanocytic proliferation.

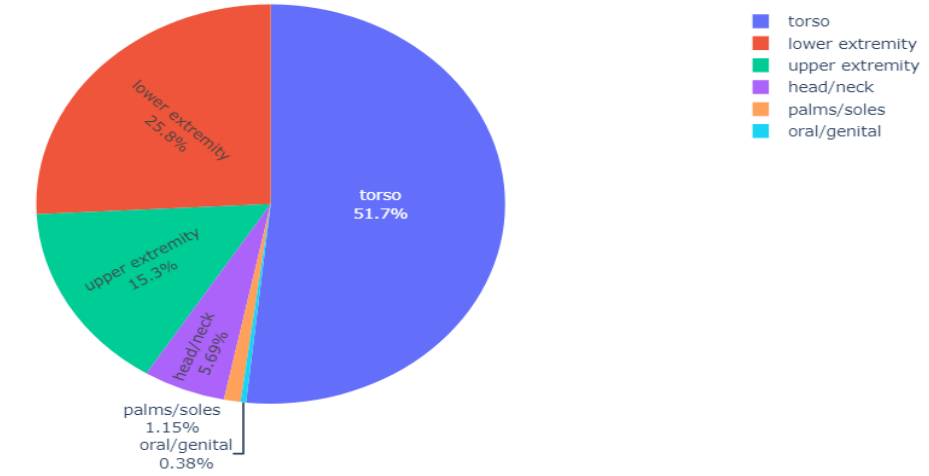

Figure 6 ratio of location of images.

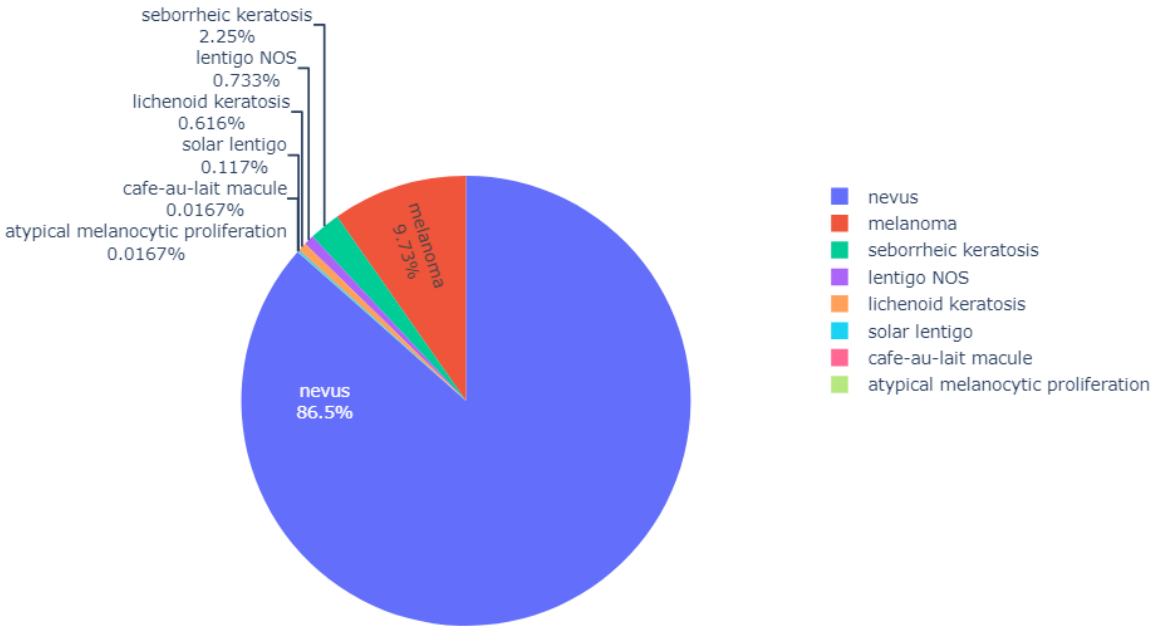

Figure 7. ratio of diaganosis

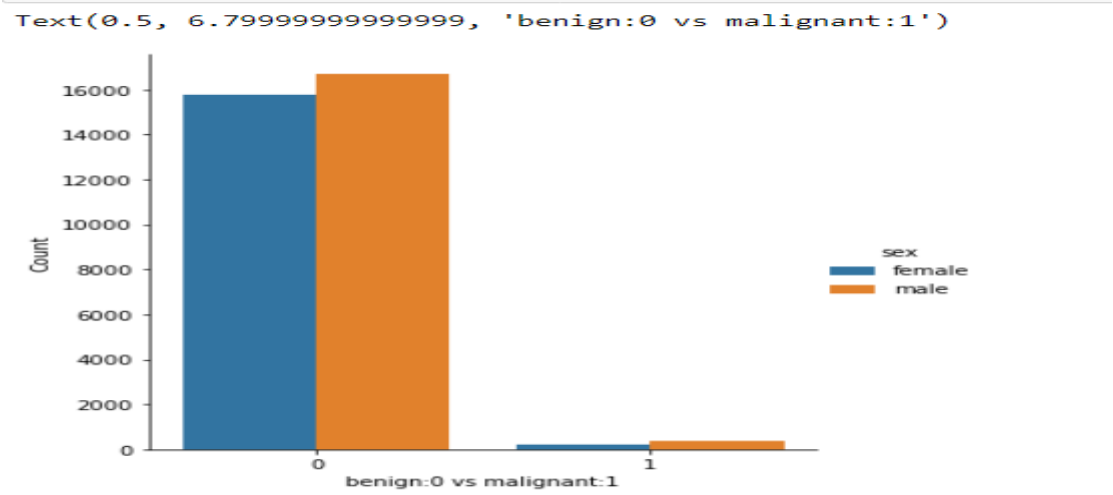

Figure 8. ratio of sex with respected to benign and malignant.

The above figure8 show the ratio male and female present in the data. The figure shows that ratio of female patients are high as compaired to male in both benign and malignant cases. However the difference very few. Also from the figure we find that benin image of vey high compaired to malignant.

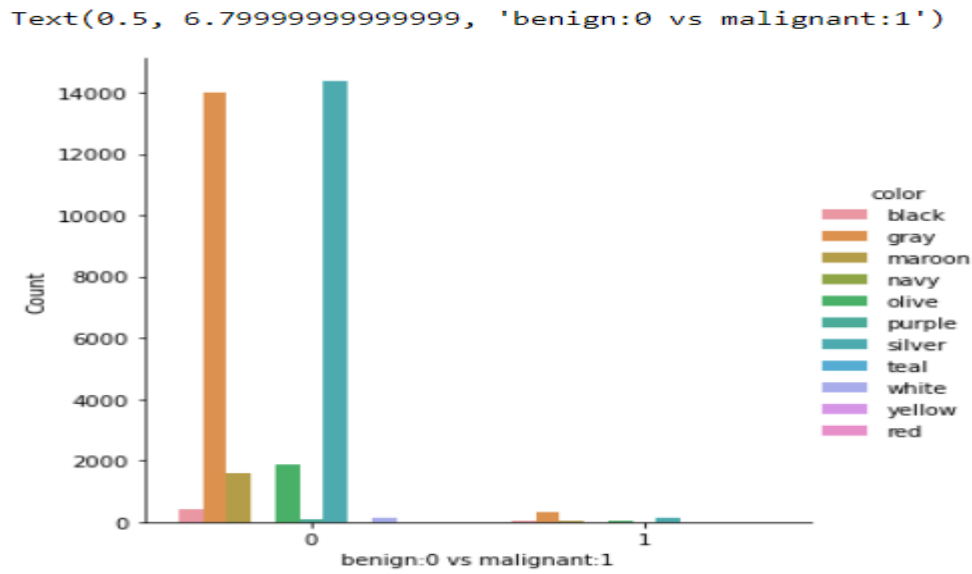

Figure 9. ratio of image color

From th above figure 9 we plot the bar graph of color extracted from each images. It is clear that majority of the images are in siliver and gray color.this color only shows that average color that present in the image.

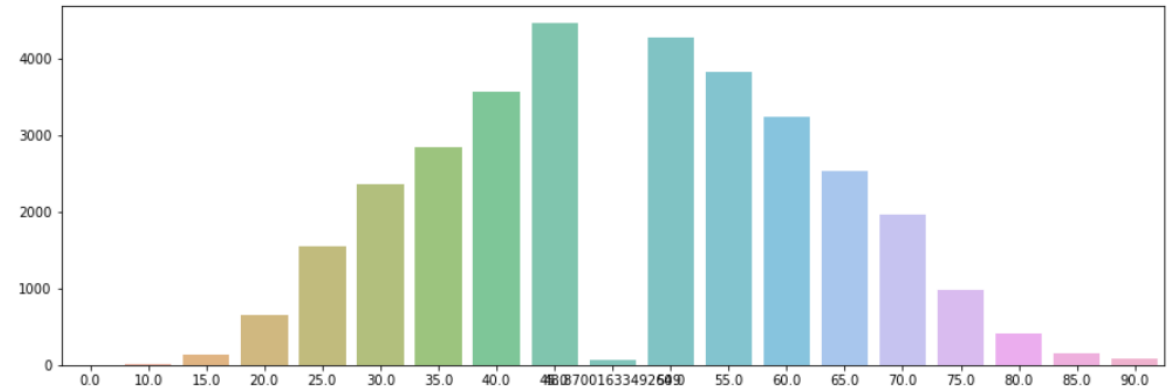

Figure 10. ratio of age of paients

The figure 10 shows the bar chart of ratio of average age of patients in our dataframe. The highest number of patients are coming in the age group of 40 to 45 and 50 to 55. So the disease mainly affected the middel aged persons. Kids and old peoples are very rere in the study.

4 Design Specification

We used three machine-learning algorithms in our research: logistic regression, SVM, and Random Forest. All three approaches are employed in regression analysis to assess connections between a 'target' variable and the other variables, generally known as independent variables. It's utilized to determine the strength of a relationship between two variables and how they'll interact.

4.1 Random Forest

The supervised learning approach is used by Random Forest, a very popular and highly used machine learning technique. Breiman (2001) developed a random forest variant of the random decision tree technique. It can be used for both classification and regression problems in learning algorithms**.** When tree length increases, the problem of overfitting and excessive variance is reduced by using random forests. This approach involves selecting a subset of records and columns, and then fitting a decision tree to each bundle. Instead of extracting patterns just from strong variables, bagging gives a model for extracting patterns from all characteristics. Random forest determines the mean of all decision tree projected values while testing. It requires fewer time for training compared with other methods. It accurately simulates output and operates rapidly, even with a large dataset..

4.2 Logistic Regression

Logistic regression is a frequent categorisation method. It belongs to the linear classifier category, along with polynomial as well as linear regression. The findings of logistic regression are simple to comprehend and are quick to compute. Although it is essentially a binary classification algorithm, it may also be applied to problems involving several

categories. To convert anticipated values to chances, the Sigmoid function is utilized. Any actual number is converted to a number between 0 and 1. The sigmoid function is used in machine learning to translate forecasts to chances. Logistic regression examines the relationship seen between categorical dependent variable and one or more predictor variable by assessing components using a logistic function, which is the cumulative distribution function of logistic distribution. As a result, it uses comparable strategies to tackle the same set of issues as probit regression, with later employing a cumulative normal distribution curve instead.

4.3 Support Vector Machine

Support-vector machines are supervised teaching methods used in machine learning to analyse data for classification and regression. It was created at AT&T Bell Laboratories by Vladimir Vapnik and co-workers. One of the most accurate forecast methodologies is SVMs, which are exposed to quantitative capacity for building or VC theory. Given a set of training samples, each categorized as largely into two subgroups, an SVM training method creates a model that assigns training instances to one of two categories, composed of non binary linear classifier. SVM converts training sets into spatial coordinates to expand the gap between the two classes as much as feasible.

5 Implementation

Reading the data and producing a dataset sub-sample with only the relevant variables was the first step in our study. First we removed the patient id, image name columns from the data that are no valid in our study. Also in the column diagnosis most of the values are unknown that may affect our prediction. Created a column that contain the colours which is the colour extracted from the images. Last we change all the categorical variable to numerical variables. Because the column sex, color, anatom_site_general_challenge are in categorical variables. Which we can use in our model building. Before implementing the model we divide our dataset test and train in the ratio of 70:30.

Python version 3.8 is used for implementing the research, and the code is written
using Jupiter Notebook. Pandas and os used to read the datasets. Data was stored in the local machine. Matplotlib and ploty libraries used for the visualization. Webcolorws used to extract the colours from images. And sklearn used to split the dataset and implement the models. The machine that was utilized in the study had the following configuration:

- Intel Core i5 (windows)
- 8 Gb Ram
- 1 TB SSD

6 Evaluation

All three models have been subjected to a thorough examination. To build the best models, hyperparameter tuning was done to get the best results. After all the processing and cleaning we have the final dataset of 32534 record. Then the dataset divided intotest set and train set in the ratio of 70:30. Also convert the infinite or null values remains in the data frame using

nan_to_num() function. Then all the result obtained from each model is evaluated. It was done by using accuracy, precision, roc curve, f1 score.

6.1 Random Forest

Some cross validation was done on the model to analyse its performance. The accuracy of the model is 74% and it shows that model is not bad in prediction. The classification report shows gives information about the precision, recall and f1 score. Precision is about .72 which is also nice value. Recall and f1-score are .76 and .74 respectively. All these result are shown in the below figure 11. And in the roc curve the auc value is 0.82.

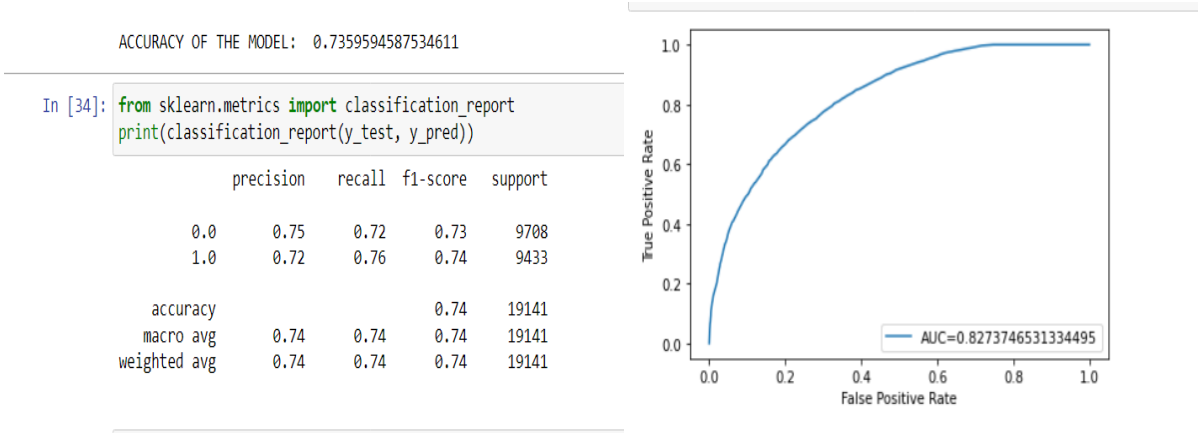

Figure 11. roc curve and classification report of random forest

6.2 Support Vector Machine

The support vector machine followed the same process. The model has a 0.63 accuracy. We examined the accuracy, recall, and f1-score from the classification report. Which are, correspondingly, 0.66, 0.56, and 0.61. Which is clearly shown in the figure 12. Given below. From the roc curve we got the area under ROC curve as 0.75. From these values we can say that Support vector machine is an average model for this prediction.

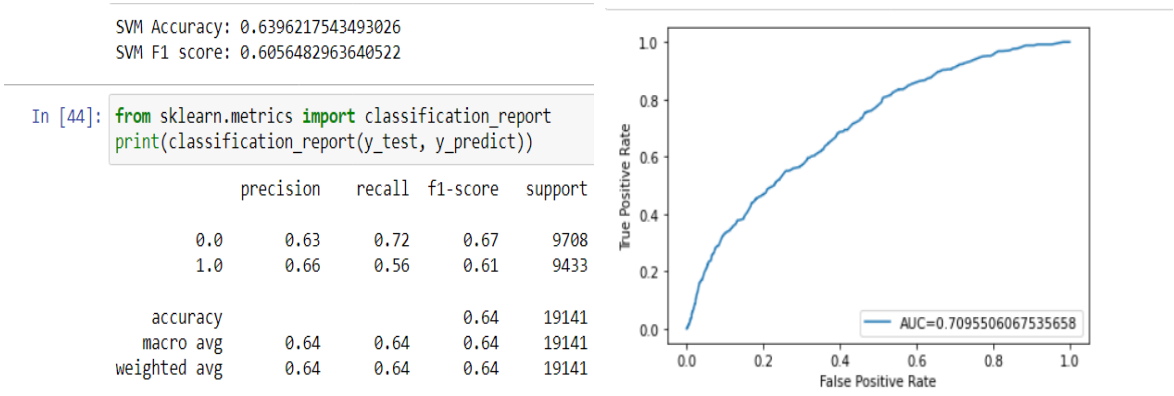

Figure 12 Classification report and ROC curve of SVM

6.3 Logistic Regression

The roc curve, accuracy, precision, recall, and f1 score are also used to evaluate the performance of the model in logistic regression. Figure 13 provides us with the values. The model's accuracy is 0.65, as seen by this graph. A accuracy of 0.64 has been determined. Both of the recall and f1 score is 0.65. from roc curve we can find that the area under Roc curve is 0.65. from this we can conclude that the logistic regression model is an average algorithm for this prediction.

```
In [39]: y_pred = logreg.predict(X_test)
         print('Accuracy of logistic regression classifier on test set: {:.
```

```
Accuracy of logistic regression classifier on test set: 0.65
```

```
In [40]: from sklearn.metrics import classification_report
         print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

         0.0       0.66      0.65      0.65      9708
         1.0       0.64      0.65      0.65      9433

    accuracy                           0.65     19141
   macro avg       0.65      0.65      0.65     19141
weighted avg       0.65      0.65      0.65     19141
```

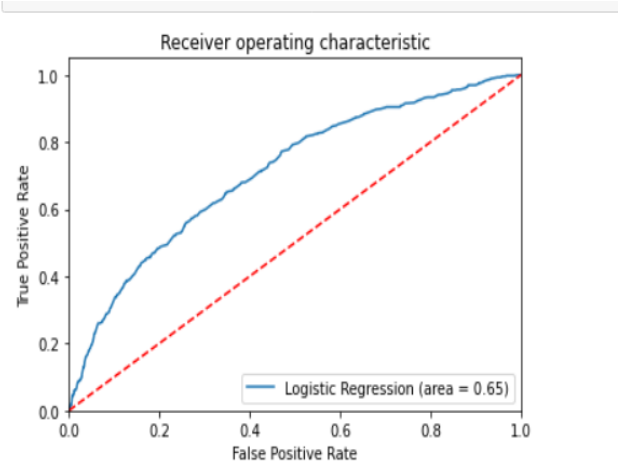

Figure 13. clasification report and roc curve of Logistic regression

6.4 Discussion

After applying the model to the dataset, we determined that the random forest classifier is the most accurate among the other support vector machine and logistic regression models. The random forest classifier predicts the output with a 74 percent accuracy. On the other hand, SVM and Logistic regression have only mediocre prediction accuracy. Based on all other assessment techniques, such as accuracy, f1-score, recall, and ROC curve, the random forest classifier is also the best among these models. These values are commonly used in classification models to assess their performance. But we can't say that random forest classifier is the best model for this prediction of malign melanoma.

There are many other classification model which may make more accuracy in the prediction. And our dataset is highly unbalanced that also make some difficulty in the better performing of the model. We can use the dataset after modifying it and make more accurate prediction.

7 Conclusion and Future Work

The major goal of our research is to use several machine learning approaches to discover malignant and benign melanoma and to determine the best model among them. Unlike prior studies, we employed picture and tabular data to make our predictions. We identified features from the photos and appended them to a tabular that had different patient information, such

as sex, age, diagnosis, and so on. Using these images, I have extracted the colour from the photos and attached it to the data frame. After analysing the features in the data I found that. Sex does not have a role in the development of this malignancy. The condition primarily affects people in their forties and fifties. It is also found in the torso area for the vast majority of humans.

I have applied 3 different machine learning algorithms like Support vector machine, Random forest, and Logistic Regression. From this all, random forest shows the best performance. It shows an accuracy of 74% and all others are 67 and 65 percentage. Logistic regression and SVM shows only average performance. There is also some limitations in the study. The dataset used in this research is highly unbalanced. That affected the performance. For further, We can make a good prediction by using a different best dataset. Also, only colour of the images is extracted from the images. We can make the study better by extracting more features from the images. Along with that using better machine learning model that make more accuracy is also we can consider for the further study.

References

Abraham, J., Heimberger, A. B., Marshall, J., Heath, E., Drabick, J., Helmstetter, A., Xiu, J., Magee, D., Stafford, P., Nabhan, C. et al. (2021). Machine learning analysis using 77,044 genomic and transcriptomic profiles to accurately predict tumor type, *Translational oncology* 14(3)

Ahmad, L. G., Eshlaghy, A., Poorebrahimi, A., Ebrahimi, M., Razavi, A. et al. (2013). Using three machine learning techniques for predicting breast cancer recurrence, *J Health Med Inform* 4(124)

Grzymala-Busse, P., Grzymala-Busse, J. W. and Hippe, Z. S. (2001). Melanoma prediction using data mining system lers, *25th Annual International Computer Software and Applications Conference*. COMPSAC 2001, IEEE, pp. 615–620

Harbour, J. W. (2014). A prognostic test to predict the risk of metastasis in uveal melanoma based on a 15-gene expression profile, *Molecular Diagnostics for Melanoma, Springer*, pp. 427–440

Johannet, P., Coudray, N., Donnelly, D. M., Jour, G., Illa-Bochaca, I., Xia, Y., Johnson, D. B., Wheless, L., Patrinely, J. R., Nomikou, S. et al. (2021). Using machine learning algorithms to predict immunotherapy response in patients with advanced melanoma, *Clinical Cancer Research* 27(1): 131–140

Kawahara, D., Tang, X., Lee, C. K., Nagata, Y. and Watanabe, Y. (2020). Predicting the local response of metastatic brain tumor to gamma knife radiosurgery by radiomics with a machine learning method, *Frontiers in Oncology* 10

Kniep, H. C., Madesta, F., Schneider, T., Hanning, U., Sch¨onfeld, M. H., Sch¨on, G., Fiehler, J., Gauer, T., Werner, R. and Gellissen, S. (2019). Radiomics of brain mri: utility in prediction of metastatic tumor type, *Radiology* 290(2): 479–487

Lynch, C. M., Abdollahi, B., Fuqua, J. D., Alexandra, R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H. and Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques, *International journal of medical informatics* 108: 1–8

Mancuso, F., Lage, S., Rasero, J., D´ıaz-Ram´on, J. L., Apraiz, A., P´erez-Yarza, G., Ezkurra, P. A., Penas, C., S´anchez-Diez, A., Garc´ıa-Vazquez, M. D. et al. (2020). Serum markers improve current prediction of metastasis development in early-stage melanoma patients: a machine learning-based study, *Molecular oncology* 14(8): 1705–1718.

Mccarthy, J. F., Marx, K. A., Hoffman, P. E., Gee, A. G., O'neil, P., Ujwal, M. L. and Hotchkiss, J. (2004). Applications of machine learning and high-dimensional visualiza tion in cancer detection, *diagnosis, and management, Annals of the New York Academy of Sciences* 1020(1): 239–262.

Mocellin, S., Ambrosi, A., Montesco, M. C., Foletto, M., Zavagno, G., Nitti, D., Lise, M. and Rossi, C. R. (2006). Support vector machine learning model for the prediction of sentinel node status in patients with cutaneous melanoma, *Annals of surgical oncology* 13(8): 1113–1122.

Shoombuatong, W., Schaduangrat, N., Pratiwi, R. and Nantasenamat, C. (2019). Thpep: A machine learning-based approach for predicting tumor homing peptides, *Computational biology and chemistry* 80: 441–451.

Thomsen, K., Iversen, L., Titlestad, T. L. and Winther, O. (2020). Systematic review of machine learning for diagnosis and prognosis in dermatology, *Journal of Dermatological Treatmen*t 31(5): 496–510.

Tseng, C.-J., Lu, C.-J., Chang, C.-C. and Chen, G.-D. (2014). Application of machine learning to predict the recurrence-proneness for cervical cancer, *Neural Computing and Applications* 24(6): 1311–1316.

Tsur, N., Kogan, Y., Avizov-Khodak, E., Vaeth, D., Vogler, N., Utikal, J., Lotem, M. and Agur, Z. (2019). Predicting response to pembrolizumab in metastatic melanoma by a new personalization algorithm, *Journal of translational medicine* 17(1): 1–15

Vaquero-Garcia, J., Lalonde, E., Ewens, K. G., Ebrahimzadeh, J., Richard-Yutz, J., Shields, C. L., Barrera, A., Green, C. J., Barash, Y. and Ganguly, A. (2017). Primeum: a model for predicting risk of metastasis in uveal melanoma, Investigative ophthalmology & visual science 58(10): 4096–4105

Vijayalakshmi, M. (2019). Melanoma skin cancer detection using image processing and machine learning, *International Journal of Trend in Scientific Research and Development* (IJTSRD) 3(4): 780–784

Vuong, K., McGeechan, K., Armstrong, B. K. and Cust, A. E. (2014). Risk prediction models for incident primary cutaneous melanoma: a systematic review, JAMA dermatology 150(4): 434–444.

Waheed, Z., Waheed, A., Zafar, M. and Riaz, F. (2017). An efficient machine learning

approach for the detection of melanoma using dermoscopic images, *2017 International Conference on Communication, Computing and Digital Systems* (C-CODE), IEEE, pp. 316–319