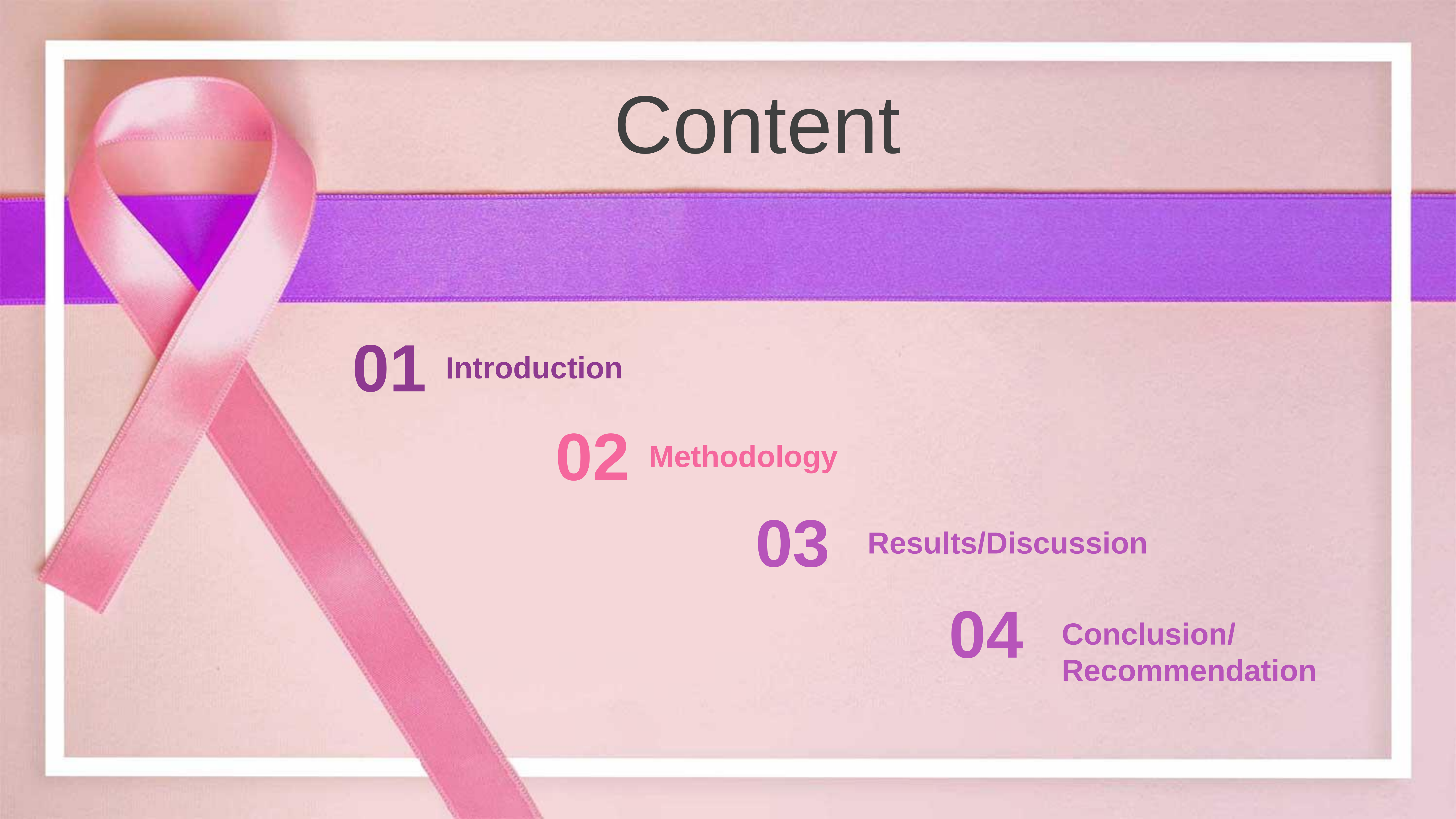
School of Computing, Engineering, and Intelligent Systems.
University of Ulster, Magee Campus
Derry, Northern Ireland

Predictive Modelling of Breast Cancer Relapse Using Machine Learning and Deep Learning Techniques

Ifedolapo Olaofe
B000



Content

01 Introduction

02 Methodology

03 Results/Discussion

04 Conclusion/ Recommendation

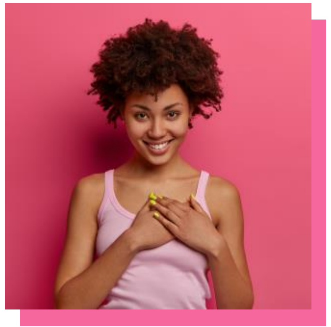
Introduction

Breast cancer is the predominant form of cancer in women and is the leading cause of cancer-related deaths among women, highlighting its significance in public health [1].

The conventional methods for cancer management rely on the "gold standard" approach, which involves three tests: clinical evaluation, radiological imaging, and pathology testing [4]. Nevertheless, this approach may be inaccurate and omits the various molecular and biological components that impact the spread of the disease.

Due to the emergence of big data analytics and machine learning techniques, there is an increasing possibility to utilise extensive patient data for creating predictive models that can accurately identify individuals with a high likelihood of experiencing cancer relapse.

The objective of this research is to create and assess predictive models for cancer relapse by utilising machine learning and deep learning methodologies.



01 The Dataset

The dataset explored for this study targeted sequencing of 2509 primary breast tumours with 548 matched normals. The dataset has 2,509 rows and 39 columns. The columns contain various clinical, pathological, and molecular characteristics of breast cancer patients, which can be used for predictive modelling and analysis. The target variable is 'relapse-free status', which indicates whether the patient experienced relapse during the study period. The dataset was spooled from an online repository provided by cBioPortal for cancer genomics [11]. Some of the columns included are

02 Existing Work

A study offered a 5-year forecast for the relapse of breast cancer. The clinicopathologic parameters of 579 individuals with breast cancer (with a recurrence prevalence of 19.3%) were analysed using a Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR), a Bagged Decision Tree classifier, to determine the likelihood of cancer relapse [9. The HPBCR demonstrated superior performance to the other classifiers examined, achieving a minimum sensitivity, specificity, precision, and accuracy of 77%, 93%, 95%, and 85%, respectively [9].

A novel support vector machine (SVM)--based prognostic model was developed to predict breast cancer recurrence within 5 years of surgery in the Korean population and compared to existing models [10]. The SVM-based 'breast cancer recurrence prediction based on SVM (BCRSVM)' outperformed other prognostic models (area under the curve=0.85, 0.71, 0.70 for BCRSVM, Adjuvant! Online, and NPI).

Methodology

01 Data Cleaning & Treating Missing Values

- Columns containing missing values exceeding 20% were eliminated.
- Additional columns such as 'patient_id', 'study_id', and 'sample_id' that are irrelevant were removed.
- To achieve 'relapse-free status', the 21 records that do not contain these values were excluded.
- Columns that contain only one level of data representation, such as 'cancer type', 'number of samples per patient', 'sample type', and 'sex' were eliminated.
- The data contained 2,488 records and 14 columns after the data cleaning process.

02 Exploratory Data Analysis

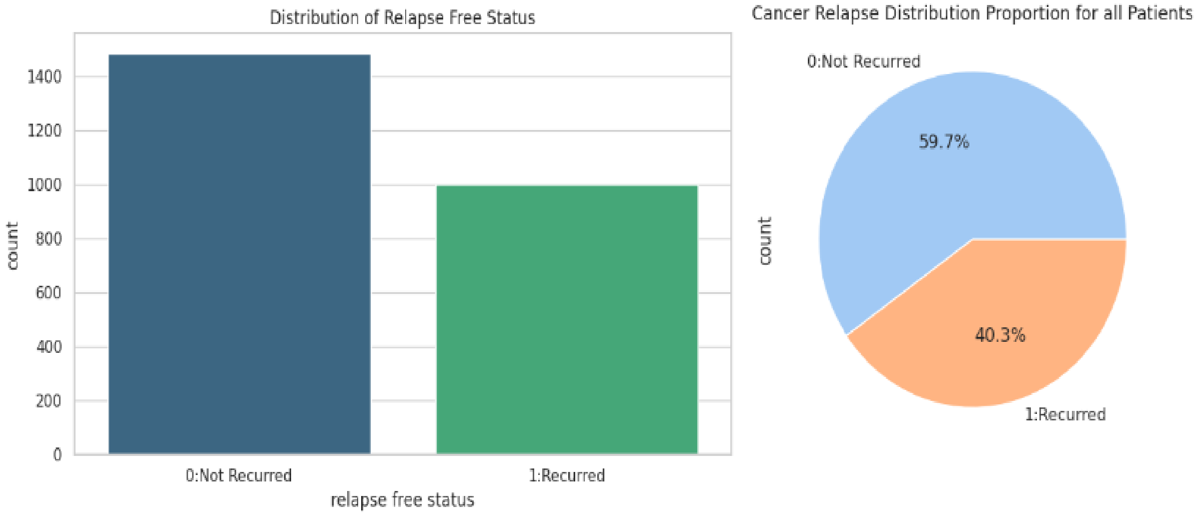


Figure 1: Distribution of cancer relapse

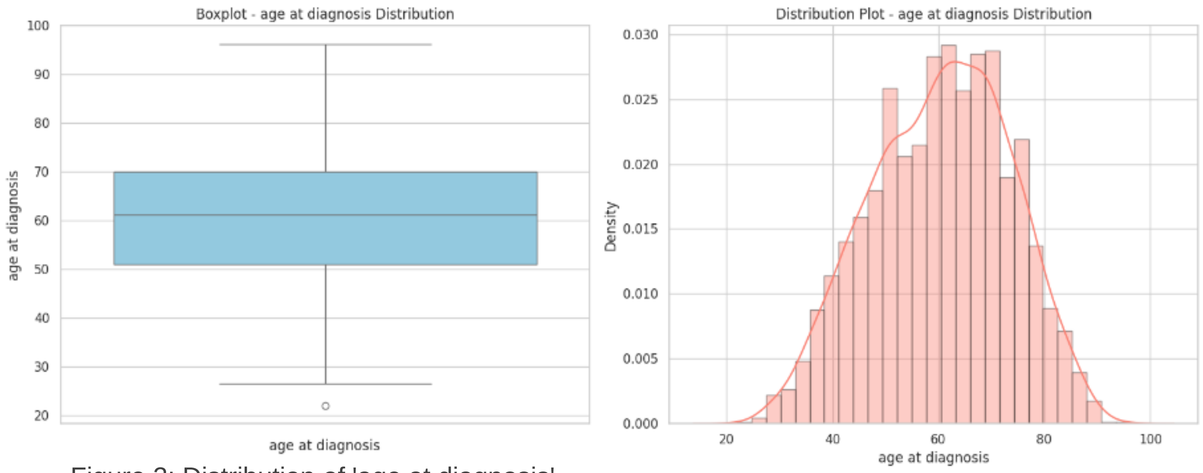


Figure 2: Distribution of 'age at diagnosis'

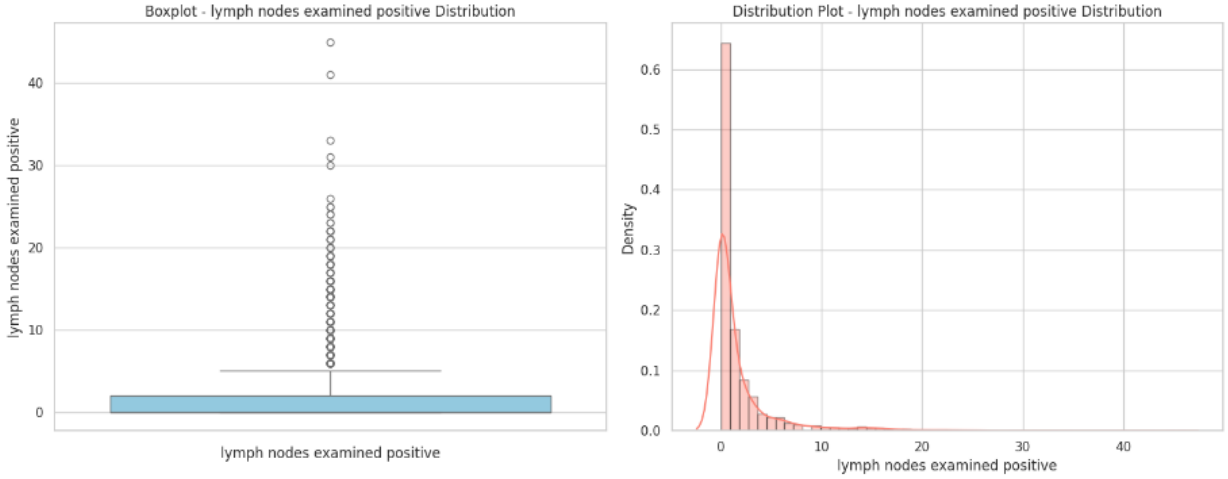


Figure 3: Distribution of 'lymph nodes examined positive'

02 Exploratory Data Analysis

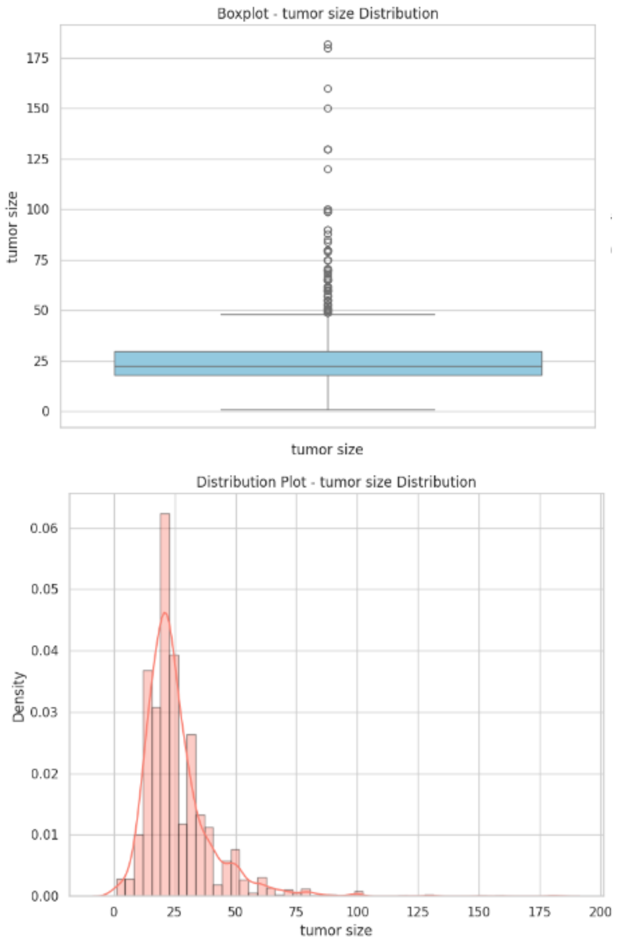


Figure 4: Distribution of 'tumour size'

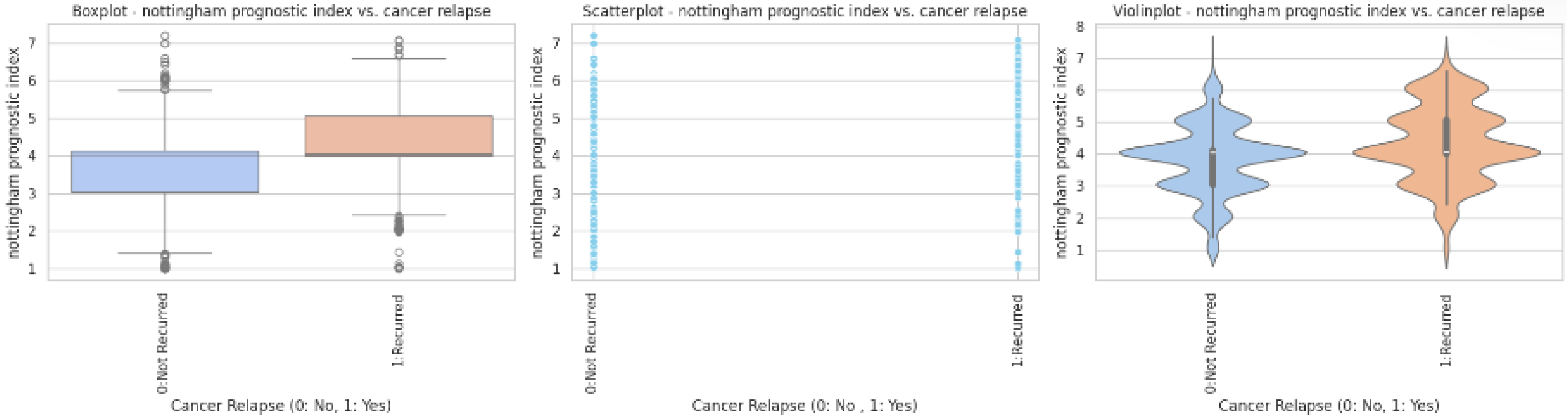


Figure 5: 'Nottingham prognostic index' vs cancer relapse

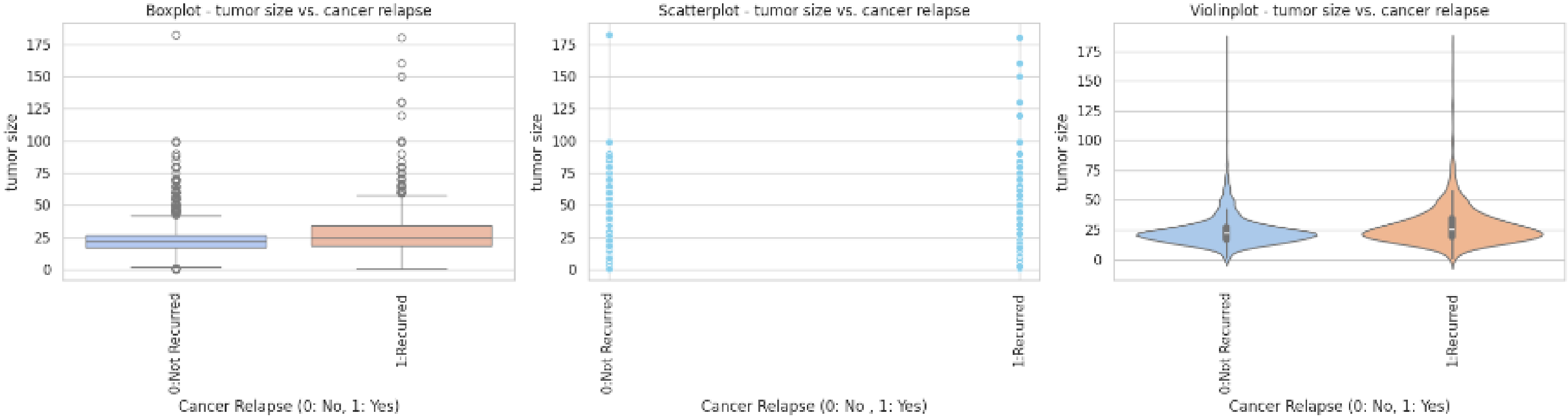


Figure 6: 'tumour size' vs cancer relapse

03 Data Preprocessing

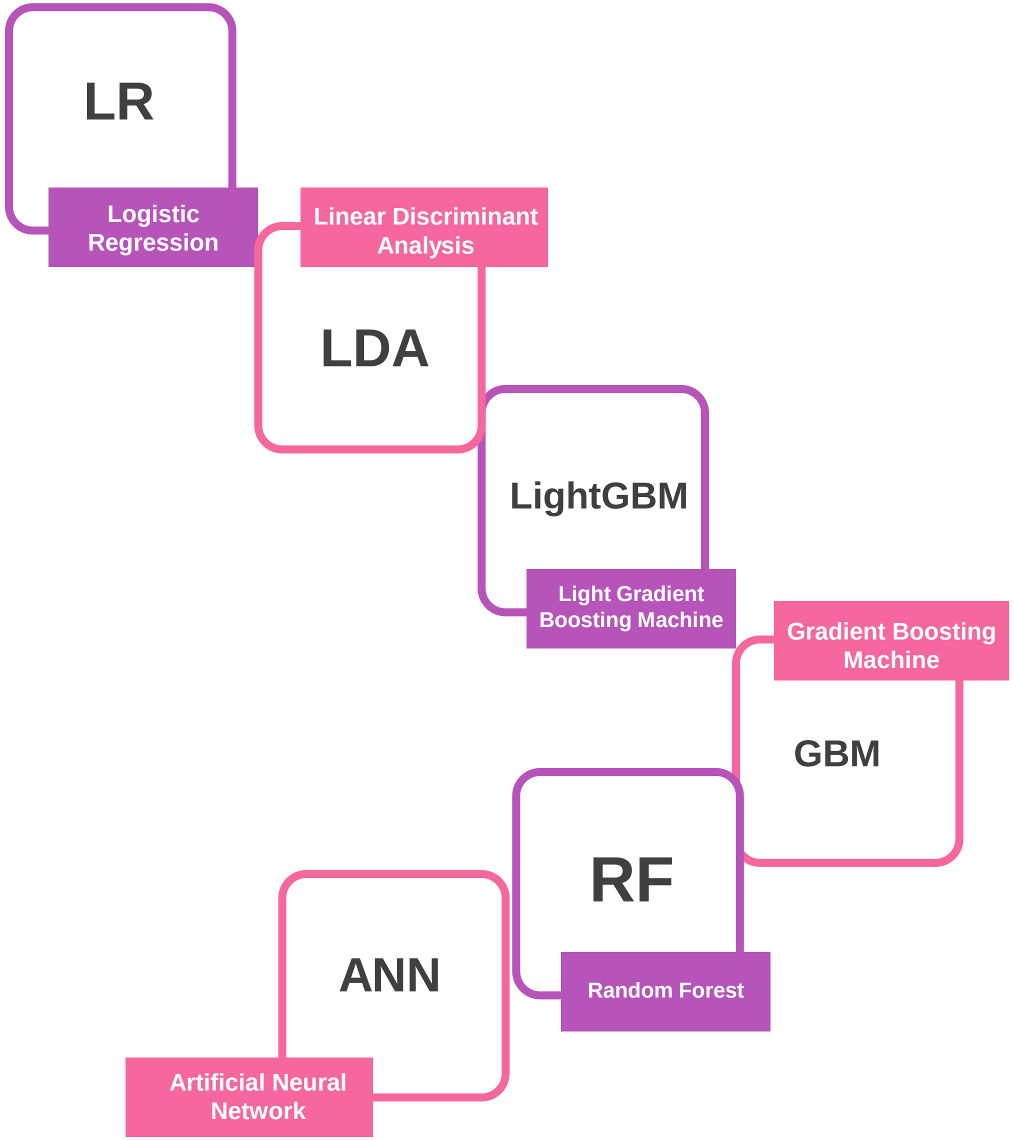
Encoded value	Original value
1	Breast
2	Breast Angiosarcoma
3	Breast Invasive Ductal Carcinoma
4	Breast Invasive Lobular Carcinoma
5	Breast Invasive Mixed Mucinous Carcinoma
6	Breast Mixed Ductal and Lobular Carcinoma
7	Metaplastic Breast Cancer

Table 1: Encoded values for 'cancer type detailed'.

04 Feature Selection

The data was divided into training and testing sets in a 70:30 ratio. The RFE was initialised using the Logistic Regression model. The initial number of features was 25, and after applying RFE, only 20 features were chosen for modelling.

05 Model Training



LR

Logistic Regression

Linear Discriminant Analysis

LDA

LightGBM

Light Gradient Boosting Machine

Gradient Boosting Machine

GBM

RF

Random Forest

ANN

Artificial Neural Network

Results

Model	Dataset	Accuracy	Precision	Recall	F1
LR	Train	0.645	0.649	0.645	0.599
	Test	0.661	0.654	0.661	0.625
LDA	Train	0.644	0.650	0.644	0.597
	Test	0.664	0.658	0.664	0.627
LightGBM	Train	0.663	0.673	0.663	0.627
	Test	0.664	0.656	0.664	0.632
GBM	Train	0.658	0.673	0.658	0.613
	Test	0.676	0.677	0.676	0.639
RF	Train	0.662	0.688	0.662	0.613
	Test	0.673	0.675	0.673	0.634
XGBoost	Train	0.696	0.739	0.696	0.655
	Test	0.674	0.678	0.674	0.634
ANN	Train	0.643	0.663	0.643	0.585
	Test	0.665	0.663	0.665	0.624

Table 2: Evaluation of model performance

- On the training set, the logistics regression model exhibits an accuracy of 64.45%, a precision of 64.96%, a recall of 64.45%, and an f1 score of 59.98%.
- The Logistic Regression model and LDA perform similarly, but LDA has slightly higher accuracy and F1 score.
- LightGBM outperforms LDA and LR with a test accuracy score of 66.4% and an F1 score of 63.22%.
- GBM performs slightly better than other models.
- Other than XGBoost and Gradient Boosting Machine, RF performs slightly better than LGBM.
- XGBoost performs closest to GBM.
- ANN has the lowest F1 score but is slightly more accurate than LR, LDA, and LGBM.

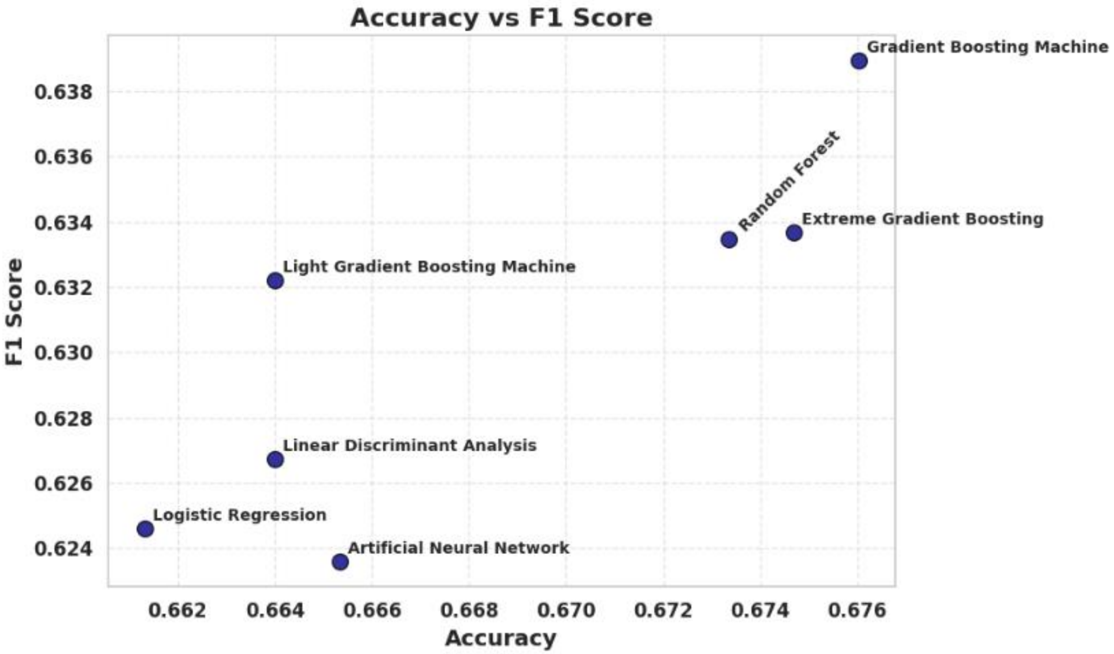


Figure 7: Accuracy vs F1 scores across Models

Fig. 7 shows that the:

- Gradient Boosting Machine Model had the best performance.
- LR had the lowest performance based on accuracy (66.1%).
- ANN had the lowest performance based on f1 score (62.4%)

Discussion

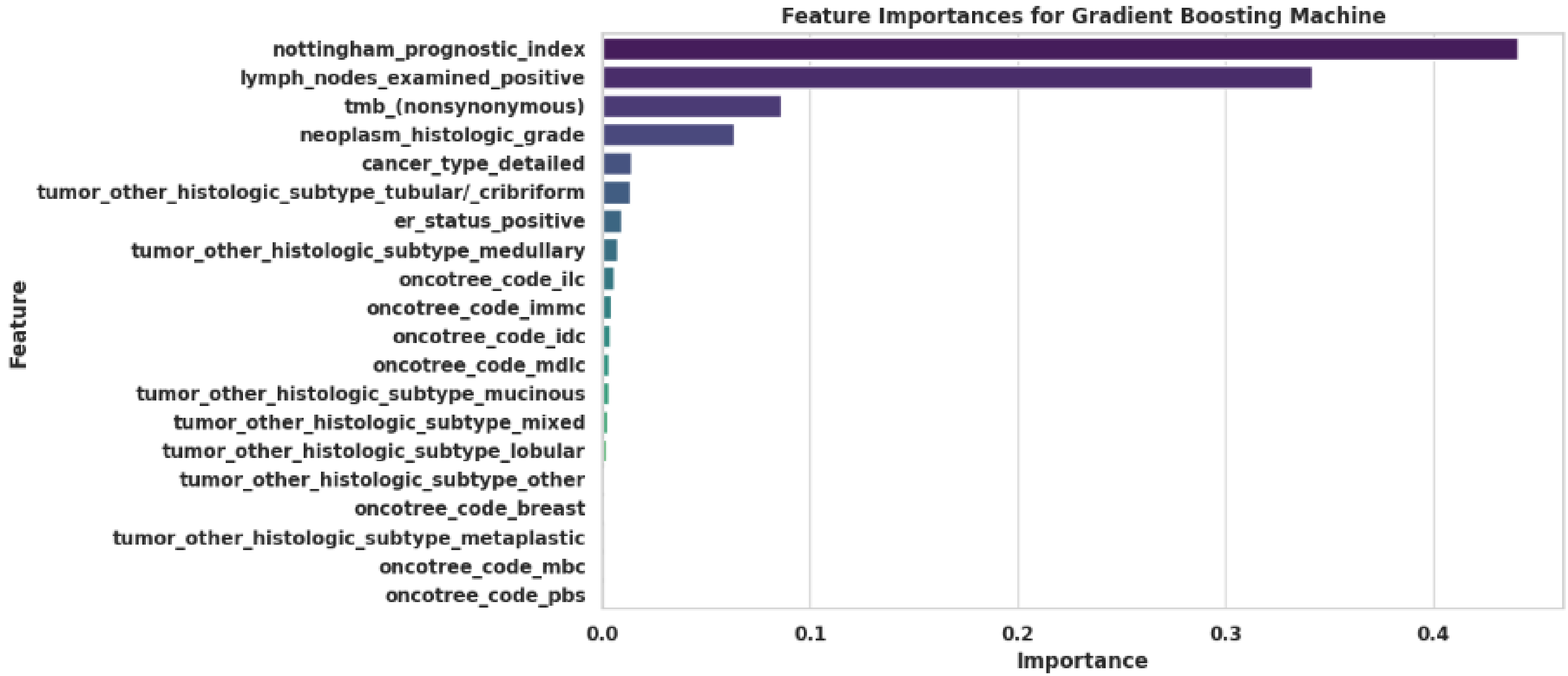


Figure 10: Feature importance for Gradient Boosting machine.

Conclusion

The key features that have been identified, particularly the Nottingham Prognostic Index, the number of positive lymph nodes examined, and the tumour mutational burden (TMB), are highly significant in predicting the likelihood of relapse. Their importance indicates that both the characteristics of the tumour and factors related to the patient significantly contribute to the probability of relapse.

When evaluating the risk of relapse in breast cancer patients, healthcare professionals are recommended to give priority to monitoring and assessing these key factors. Moreover, by utilising these features, it is possible to identify specific characteristics that can be used to develop focused interventions and treatments. This can lead to improved patient outcomes and a decreased likelihood of relapse.

References

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2022," *Ca*, vol. 72, no. 1, pp. 7–33, Jan. 2022, doi: 10.3322/caac.21708.

[2] M. L. Ascierto *et al.*, "A signature of immune function genes associated with recurrence-free survival in breast cancer patients," *Breast Cancer Research and Treatment*, vol. 131, no. 3, pp. 871–880, Apr. 2011, doi: 10.1007/s10549-011-1470-x.

[3] D. Lindquist, S. Kvarnbrink, R. Henriksson, and H. Hedman, "LRIG and cancer prognosis," Acta Oncologica, vol. 53, no. 9, pp. 1135–1142, Sep. 2014, doi: 10.3109/0284186x.2014.953258.

[4] N. Fatima, L. Liu, H. Sha, and H. Ahmed, "Prediction of breast cancer, comparative review of machine learning techniques, and their analysis," IEEE Access, vol. 8, pp. 150360–150376, Jan. 2020, doi: 10.1109/access.2020.3016715.

[5] D. Hong *et al.*, "Epithelial-to-mesenchymal transition and cancer stem cells contribute to breast cancer heterogeneity," *Journal of Cellular Physiology*, vol. 233, no. 12, pp. 9136–9144, Jul. 2018, doi: 10.1002/jcp.26847.

[6] Y. Mitobe *et al.*, "PSF promotes ER-Positive breast cancer progression via posttranscriptional regulation of ESR1 and SCFD2," *Cancer Research*, vol. 80, no. 11, pp. 2230–2242, Jun. 2020, doi: 10.1158/0008-5472.can-19-3095.

[7] L. Sun, Y. X. Zhu, Q. Qian, and L. Tang, "Body mass index and prognosis of breast cancer," *Medicine*, vol. 97, no. 26, p. e11220, Jun. 2018, doi: 10.1097/md.0000000000011220.

[8] M. Schootman, D. B. Jeffe, W. E. Gillanders, and R. Aft, "Racial disparities in the development of breast cancer metastases among older women," *Cancer*, vol. 115, no. 4, pp. 731–740, Jan. 2009, doi: 10.1002/cncr.24087.

[9] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. À. Mañanas, and F. Mokarian, "A hybrid computer-aided-diagnosis system for Prediction of Breast Cancer Recurrence (HPBCR) using optimized ensemble learning," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 75–85, Jan. 2017, doi: 10.1016/j.csbj.2016.11.004.

[10] W. Kim *et al.*, "Development of novel breast cancer recurrence prediction model using support Vector Machine," *Journal of Breast Cancer/Journal of Breast Cancer*, vol. 15, no. 2, p. 230, Jan. 2012, doi: 10.4048/jbc.2012.15.2.230.

[11] https://www.cbioportal.org/study/clinicalData?id=brca_metabric

[12] S. Lemeshow and D. Hosmer, "*Logistic regression*", Methods and Applications of Statistics in Clinical Trials, p. 365-379, 2014. https://doi.org/10.1002/9781118596333.ch21

[13] R. Wang, "Comparison of decision tree, random forest and linear discriminant analysis models in breast cancer prediction," *Journal of Physics. Conference Series*, vol. 2386, no. 1, p. 012043, Dec. 2022, doi: 10.1088/1742-6596/2386/1/012043.

[14 K. M. M. Uddin, N. Biswas, S. T. Rikta, S. K. Dey, and A. Qazi, "XML-LightGBMDroid: A self-driven interactive mobile application utilizing explainable machine learning for breast cancer diagnosis," *Engineering Reports*, vol. 5, no. 11, May 2023, doi: 10.1002/eng2.12666.

[15] C. Yan, Q. Liu, M. Nie, W. Hu, and R. Jia, "Comprehensive analysis of the immune and prognostic implication of TRIM8 in breast cancer," *Frontiers in Genetics*, vol. 13, Mar. 2022, doi: 10.3389/fgene.2022.835540.

[16] B. M. Al-Maqaleh, A. A. Al-Mansoub, and F. N. Al-Badani, "Forecasting using Artificial Neural Network and Statistics Models," *International Journal of Education and Management Engineering*, vol. 6, no. 3, pp. 20–32, May 2016, doi: 10.5815/ijeme.2016.03.03.

[17] J. P. Monteiro, D. Ramos, D. Carneiro, F. Duarte, J. M. Fernandes, and P. Nováis, "Meta-learning and the new challenges of machine learning," *International Journal of Intelligent Systems*, vol. 36, no. 11, pp. 6240–6272, Jun. 2021, doi: 10.1002/int.22549.

[18] S. W. J. Nijman *et al.*, "Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review," *Journal of Clinical Epidemiology*, vol. 142, pp. 218–229, Feb. 2022, doi: 10.1016/j.jclinepi.2021.11.023.

Thank You