ISSN: 2752-3829 Vol. 3 No.2, (December, 2023)

Stochastic Modelling and Computational Sciences

ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR LIVER CANCER BASED ON WEIGHT OF EVIDENCE AND INFORMATION VALUE

C Jayasundari[1] and P Arumugam[2]
Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu
cjayasundari.krmmc@gmail.com[1] and sixfacemsu@gmail.com[2]

ABSTRACT

Liver cancer is a growing global health concern, requiring early detection for effect.tive treatment. This study explores the distribution of liver cancer across various features and evaluates the predictive power of categorical variables using Information Value (IV) and Weight of Evidence (WOE). The dataset consists of 1,700 records and 11 attributes. Exploratory Data Analysis (EDA) is performed using Python to critical insights that facilitate early diagnosis and risk assessment. Data visualizations, including percentage stacked charts and boxplots, were utilized to illustrate key patterns and distributions. The findings emphasize the effectiveness of Machine Learning algorithms and Data Mining techniques in liver cancer classification. These insights can contribute to improving predictive models and aiding healthcare professionals in making informed decisions.

Keywords: *Liver Cancer, Weight of Evidence, Information Value, Hypotheses Testing, Exploratory Data Analysis*

1. INTRODUCTION

Liver cancer is a significant global health issue, contributing to high morbidity and mortality rates. The liver plays a vital role in numerous physiological processes, including metabolism, detoxification, and bile production. Any disruption to its function can lead to severe health complications, making early detection and diagnosis essential. Liver cancers can stem from various factors, including excessive alcohol consumption, viral infections, genetic predisposition, and metabolic disorders. Given the increasing prevalence of liver-related conditions, effective diagnostic tools are crucial to support healthcare professionals in early detection and timely intervention.

Traditional diagnostic methods often rely on biochemical tests and clinical assessments, which may not always provide a comprehensive understanding of cancer patterns. The advancement of computational techniques, particularly in the fields of Data Mining and Machine Learning, has opened new paths for medical research and diagnostics. By analyzing large datasets, these techniques can identify hidden patterns and significant risk factors associated with liver cancer, improving predictive accuracy and clinical decision-making.

This study employs Exploratory Data Analysis (EDA) to examine liver cancer trends and determine the predictive power of various clinical and demographic features. Various statistical tests, including t-tests and chi-square tests, are used to assess relationships between features, while visualization techniques such as boxplots and percentage stacked charts help in understanding data distribution. Furthermore, WOE and IV are utilized to assess the predictive importance of categorical variables in liver cancer diagnosis.

By making use of insights and data-driven approaches gained using statistical methods, this study aims to contribute to the growing body of research on liver cancer prediction. The findings could enhance the accuracy of early diagnosis, support clinical decision-making, and pave the way for integrating Mac+++hine Learning models in future predictive frameworks.

2. REVIEW OF LITERATURE

Liver cancer is a major global health concern, with conditions like cirrhosis, hepatitis, and liver cancer being leading causes of mortality. Early prediction plays a crucial role in patient survival. Traditional diagnostic methods rely on laboratory tests and imaging, but data-driven approaches, including statistical methods and Machine Learning, have gained traction in improving predictive accuracy.

Several studies have explored Data Mining techniques to enhance liver cancer diagnosis. Baitharu & Pani (2016) conducted a comparative analysis of classifiers including J48, Naïve Bayes, Artificial Neural Networks (ANN), and others, demonstrating that decision trees and ANN models could improve predictive performance.

Another study by Arbain & Balakrishnan (2019) highlighted the impact of imbalanced datasets and compared multiple algorithms, concluding that k- Nearest Neighbours (k-NN) outperformed others with 99.79% accuracy.

Mostafa et.al., (2021) utilized statistical Machine Learning techniques to extract significant predictors from a liver cancer dataset. They implemented multiple imputations to handle missing values and applied Principal Component Analysis (PCA) for dimensionality reduction. Among Machine Learning models, Random Forest achieved the highest accuracy 98.14%, demonstrating its effectiveness in handling complex medical datasets.

A study by Kefelegn & Kamat (2018) compared classifiers like Naïve Bayes, Support Vector Machines (SVM), and decision trees, highlighting that C4.5 decision trees performed well in selecting significant features.

Razali et.al., (2020) proposed a rule-based classification model for liver cancer detection using Azure ML. Their study compared decision trees, Naïve Bayes, and SVM, showing that hybrid approaches could further enhance predictive accuracy.

While machine learning models have demonstrated promising results, gaps remain in integrating Exploratory Data Analysis techniques for feature selection prior to modeling. Many studies focus on classifier comparison but lack in-depth analysis of individual feature importance using statistical methods like Weight of Evidence. This research aims to bridge that gap by leveraging Exploratory Data Analysis to identify key predictive factors before applying Machine Learning techniques.

3. METHODOLOGY

3.1 Weight of Evidence (WOE):

Weight of Evidence (WOE) measures the predictive power of an independent variable with respect to the target variable. Initially developed for credit scoring, WOE is used to distinguish between "good" and "bad" cases, such as identifying reliable borrowers based on loan repayment history. In healthcare, for example, WOE can be applied to predict treatment outcomes, where "bad" refers to patients experiencing adverse effects, and "good" denotes those who do not.

$$WOE = \ln\left(\frac{\%\,of\ Non-Events\ in\ a\ Group}{\%\,of\ Events\ in\ a\ Group}\right) \qquad \ldots (1)$$

- % of Non-Events in a Group: The proportion of non-event outcomes within a specific group, relative to the total non-events in the dataset.
- % of Events in a Group: The proportion of event outcomes within the same group, relative to the total events in the dataset.

3.2 Interpretation of WOE Values:

A positive WOE indicates that the percentage of non-events is higher than events in a given group. A negative WOE suggests that the percentage of events exceeds non-events in that group.

Mathematical Insight: The natural logarithm of a number greater than 1 is positive, while the logarithm of a number less than 1 is negative.

3.3 Steps to compute WOE:

- For continuous variables, divide the data into 10 groups (or fewer, based on distribution).
- Determine the count of events and non-events in each group. Calculate the percentage of events and non-events within each group.
- Compute WOE by taking the natural logarithm of the ratio of non-event percentage to event percentage.

3.4 Application of WOE:

WOE transformation converts categorical or continuous variables into numerical values, making them suitable for predictive modeling. Categories with similar WOE values exhibit comparable event and non-event proportions, ensuring consistency in behavior.

3.5 Guidelines for WOE Computation:

- Each category (or bin) should contain at least 5% of the total observations.
- No bin should have zero occurrences for both events and non-events.
- WOE values must be distinct for each category; similar groups should be merged.
- WOE should follow a monotonic trend (either increasing or decreasing).
- Missing values should be handled by assigning them to a separate bin. Generally, datasets are divided into 10 to 20 bins, ensuring each contains at least 5% of cases. A smaller number of bins capture essential patterns while filtering out noise.

3.6 Handling Zero Events or Non-Events in a Bin:

If a bin has no observations, a smoothing adjustment can be applied. One approach is adding 0.5 to both event and non-event counts within the bin before computing WOE. This prevents undefined values and ensures stability in calculations.

$$Adjusted\ WOE = \ln\left(\frac{\frac{Number\ of\ Non-events\ in\ a\ group+0.5}{Total\ number\ of\ non-events}}{\frac{Number\ of\ Events\ in\ a\ group+0.5}{Total\ number\ of\ Events}}\right) \qquad \ldots (2)$$

3.7 Advantages of Using Weight of Evidence (WOE):

- It effectively manages outliers by grouping them appropriately.
- Missing values can be handled separately by assigning them to distinct bins.
- Since WOE converts categorical variables into numerical values, there is no need for one-hot encoding or dummy variables.
- WOE transformation ensures a strict linear relationship with log odds, which is often difficult to achieve using conventional transformations like logarithmic or square-root methods.

3.8 Information Value (IV):

Information Value (IV) is a widely used method for identifying the most significant variables in a predictive model. It aids in ranking features based on their predictive strength. The IV is determined using the following formula

$$IV = \sum (\%\ of\ Non-events - \%\ of\ Events) \times WOE \qquad \ldots (3)$$

Table 1: Information values in bin

Information Value	Variable Predictiveness
Less than 0.02	Not useful
0.02 to 0.1	Weak at prediction
0.1 to 0.3	Medium at prediction
0.3 to 0.5	Strong at prediction
> 0.5	Suspiciously strong at prediction

The Information Value (IV) tends to increase as the number of bins or groups for an independent variable grows. Since IV is specifically tailored for Binary Logistic Regression, it may not be the most suitable metric for other classification models.

4. ANALYSIS AND INTERPRETATION

In this section, we explore the relationships between various features and the presence of liver cancer. Through visualizations and statistical tests, we identify significant patterns and trends that contribute to cancer prediction. The analysis includes descriptive statistics, percentage distributions, and hypothesis testing to determine the predictive strength of each feature. Key variables are examined in detail using stacked bar charts, boxplots, and statistical measures such as the t- test for continuous variables and the chi-square test for categorical ones. This helps in understanding the most influential factors affecting liver cancer diagnosis. Through Hypotheses Testing, all the variables turned out to be statistically significant.

4.1 Continuous Features vs Binary Target:

Liver cancer tends to become more prevalent in middle-aged and older individuals. The lowest rates of positive diagnoses are observed in the youngest age group, whereas the highest rates are observed in individuals aged 51.0 – 63.0.

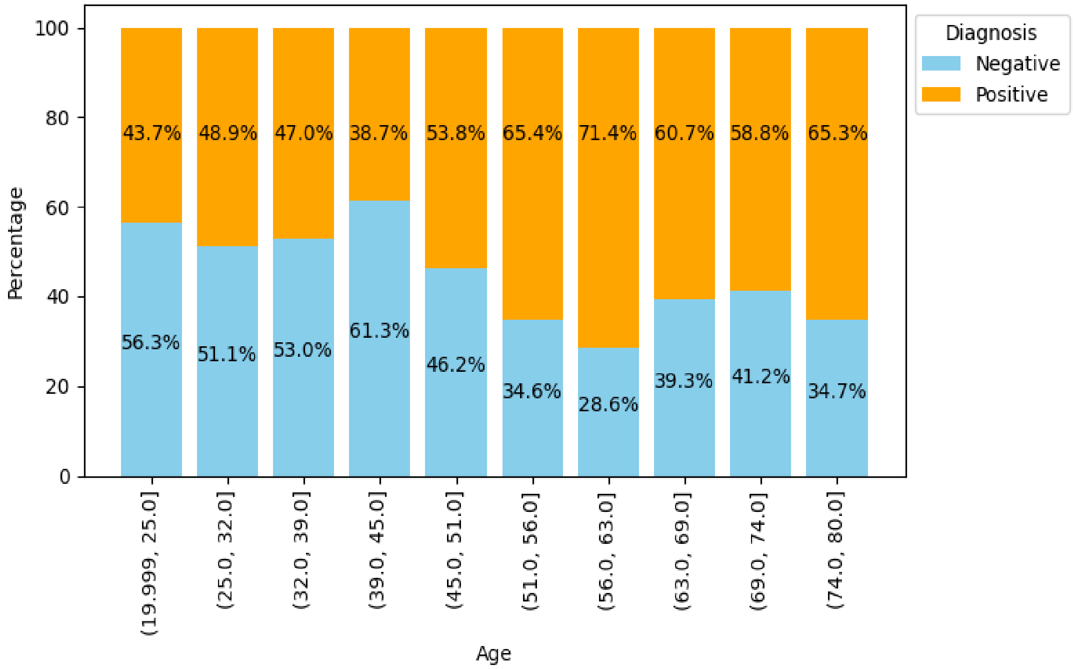

Figure 1: Distribution of diagnosis by age

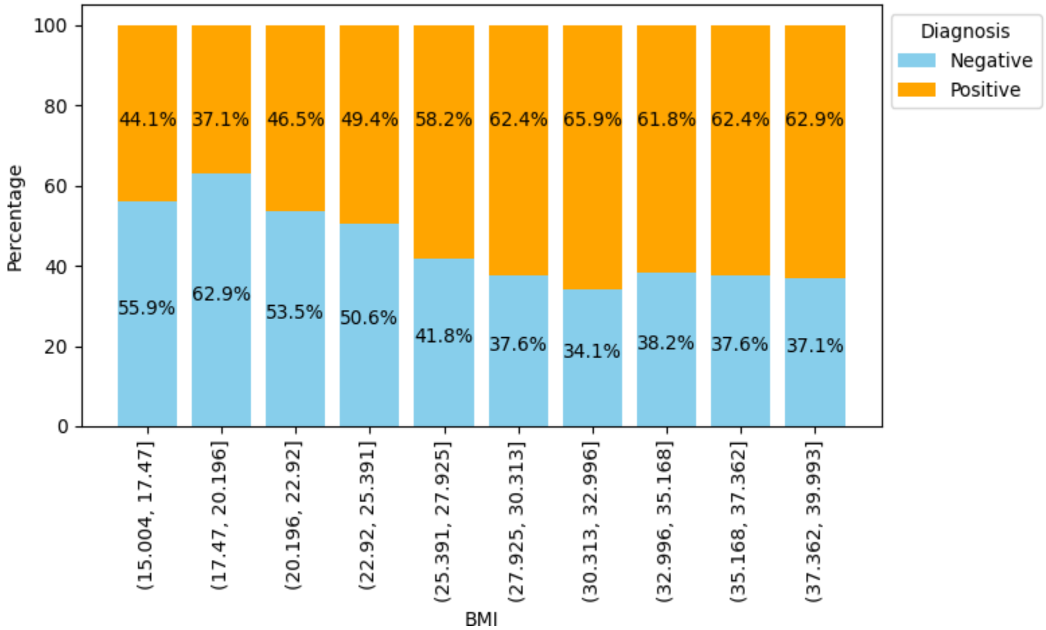

Figure 2: Distribution of diagnosis by BMI

Liver cancer prevalence increases with BMI, peaking in the overweight range (25.391 – 27.925) at 58.2% positive diagnoses. Beyond this, the proportion of positive diagnoses stabilizes around 62% – 65% in the overweight and obese categories.

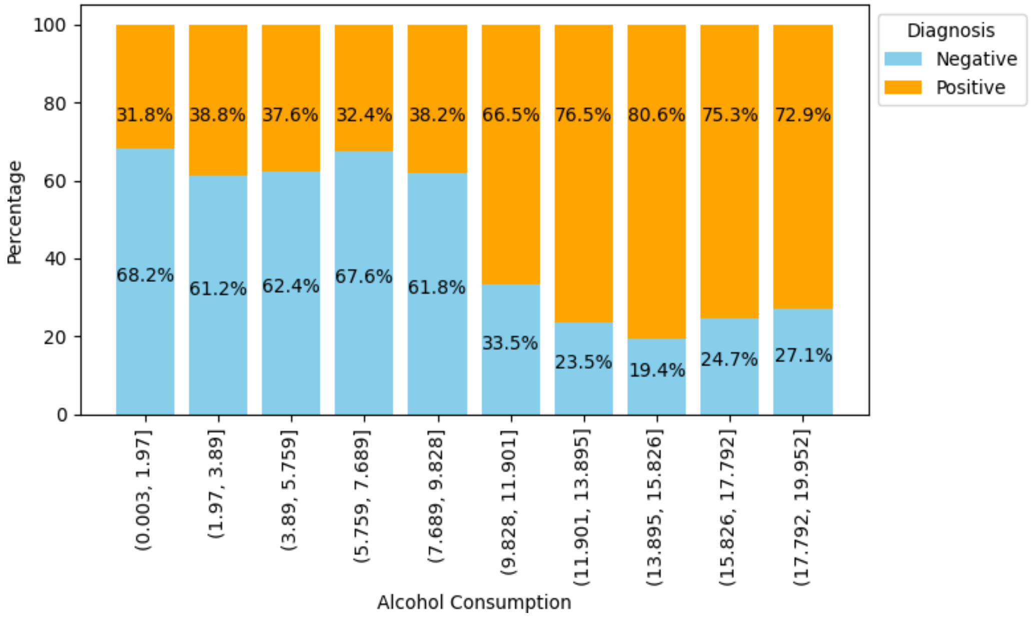

Figure 3: Distribution of diagnosis by alcohol consumption

Higher alcohol consumption strongly correlates with liver cancer diagnosis. Those with positive diagnoses had nearly double the median alcohol consumption (approx. 12.5) compared to negative cases (approx. 6.5).

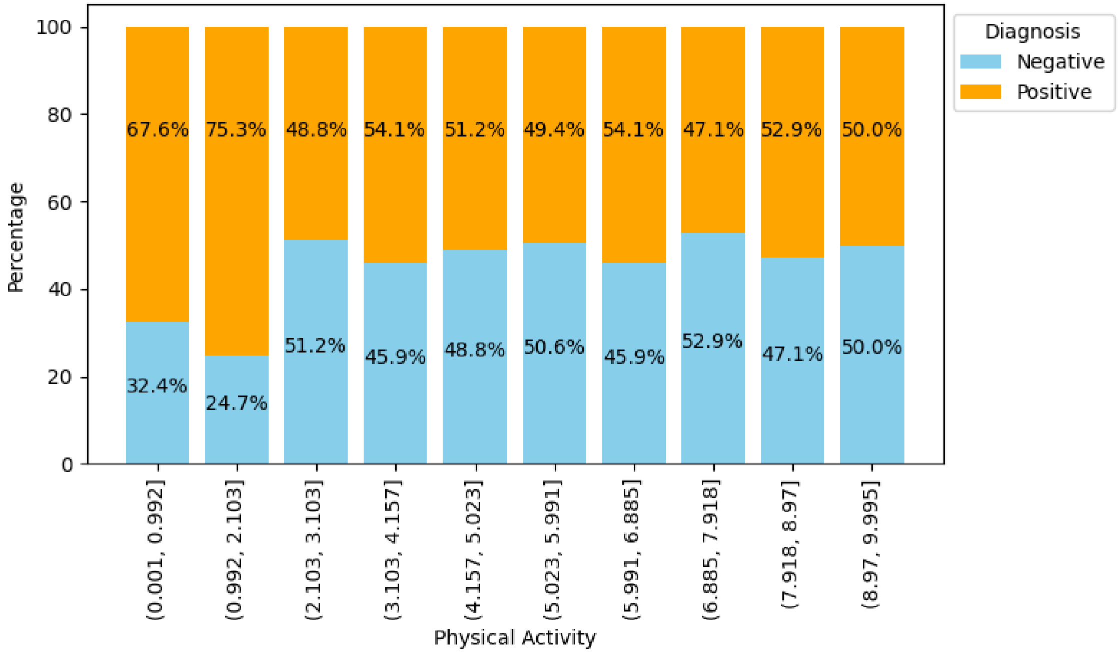

Figure 4: Distribution of Diagnosis by Physical Activity

The stacked bar chart reveals no clear trend, with diagnosis ratios remaining relatively consistent across different physical activity levels, though there's a slight indication that very low activity (0-2) may be associated with higher positive diagnoses.

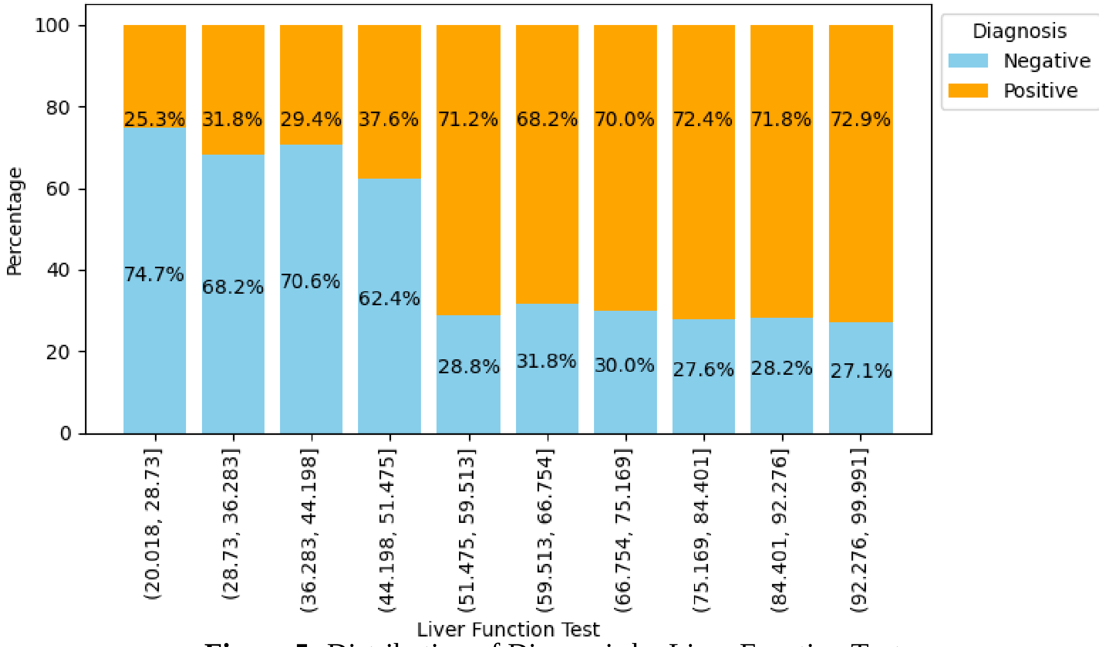

Figure 5: Distribution of Diagnosis by Liver Function Test

The stacked bar chart demonstrates a clear progression: as liver function test values increase beyond approx. 50, the proportion of positive diagnoses dramatically rises from around 30% to over 70%. This suggests that liver function test values are a strong predictor of liver cancer diagnosis.

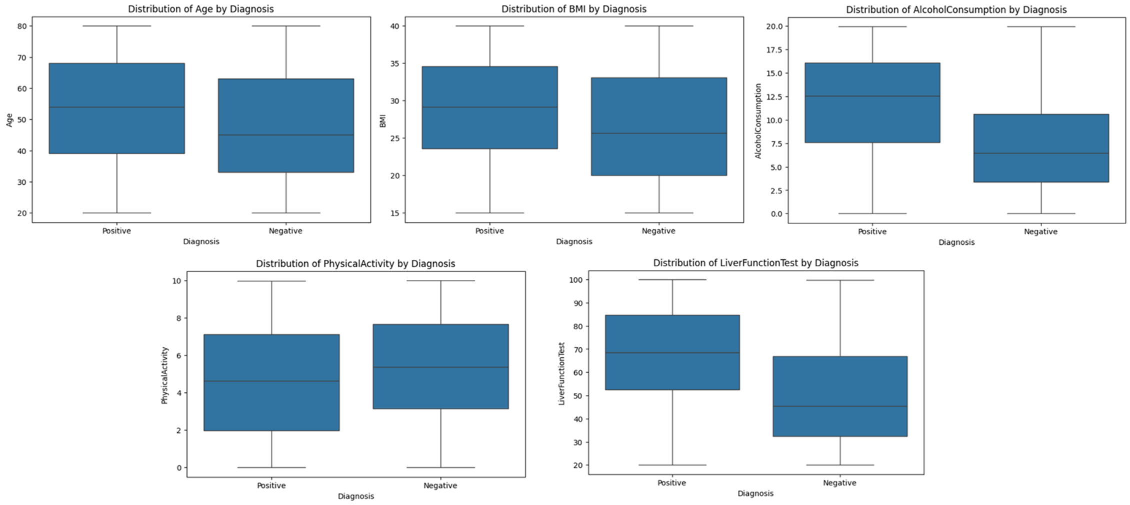

Figure 6: Boxplots for the Distribution of liver diagnosis

From the above boxplots, Age is a key risk factor, with most cases concentrated in the 50–70 range. Lower BMI correlates with fewer positive cases, making it a strong predictor. Alcohol consumption above 9–11 units significantly increases risk (>65% positive), reaching 75% at 15+ units. Physical activity shows a weak relationship, with only a slight difference between positive and negative cases. Liver function tests strongly correlate with diagnosis, with positive cases showing much higher median values (70 vs. 4).

4.2 Categorical Features vs Binary Target

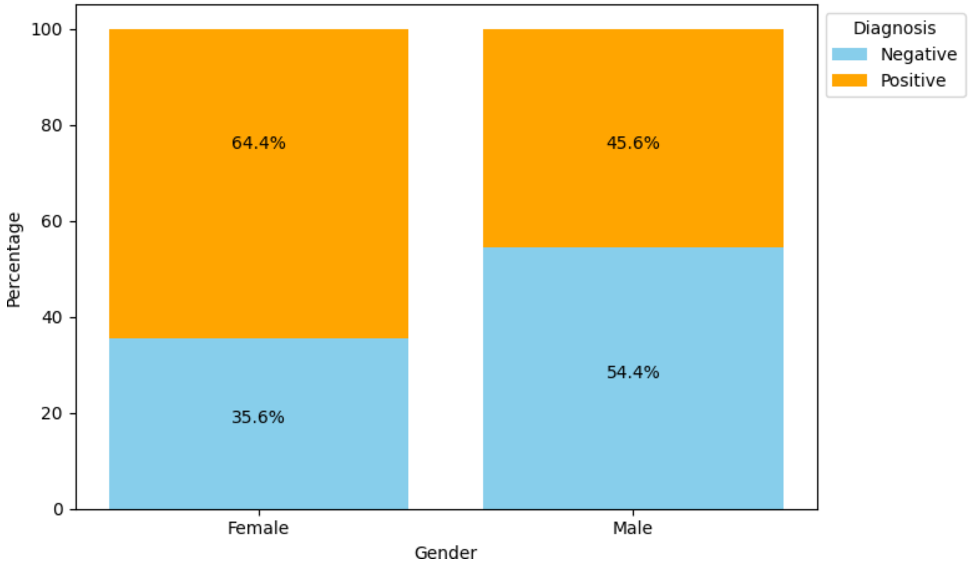

Figure 7: Distribution of Diagnosis by Gender

The data shows a gender disparity in liver cancer diagnosis. Women have a higher positive diagnosis rate (64.4%) compared to men (45.6%). Men show more negative cases (54.4%) than women (35.6%).

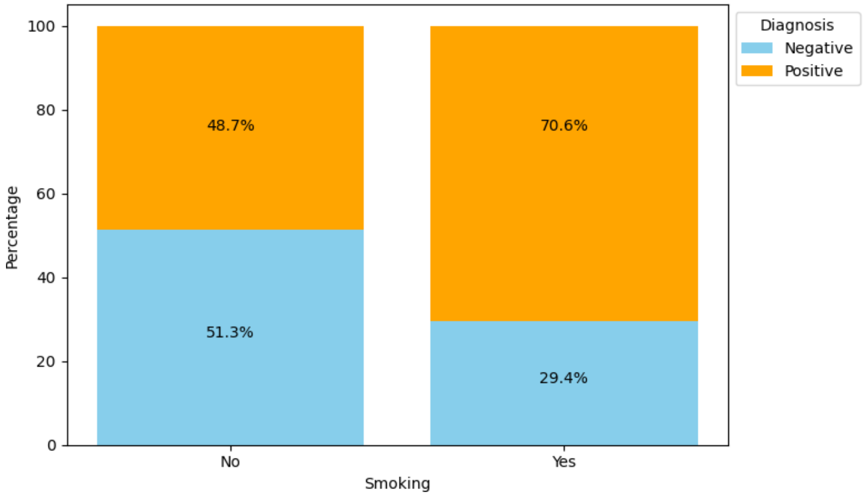

Figure 8: Distribution of Diagnosis by smoking

Smokers show significantly higher liver cancer rates (70.6% positive) compared to non-smokers (48.7% positive). Non- smokers have a more balanced distribution with slightly more negative cases (51.3%) than positive.

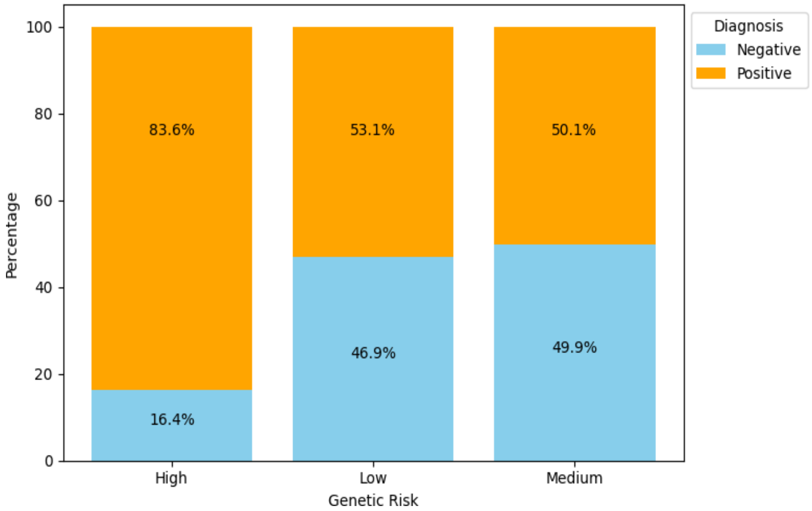

Figure 9: Distribution of Diagnosis by Genetic risk

Genetic risk is a significant factor in the likelihood of a positive diagnosis. Higher genetic risk correlates strongly with a higher percentage of positive diagnoses. Individuals with low genetic risk are much more likely to have a negative diagnosis.

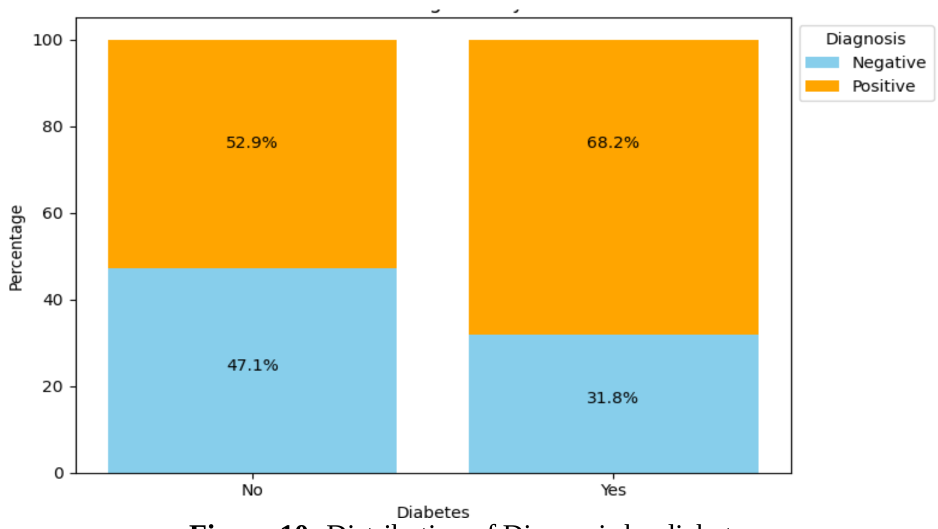

Figure 10: Distribution of Diagnosis by diabetes

There is a 15.3% increase in positive diagnoses among individuals with diabetes (68.2%) compared to those without diabetes (52.9%). This indicates that diabetes significantly elevates the risk of a positive diagnosis for liver cancer.

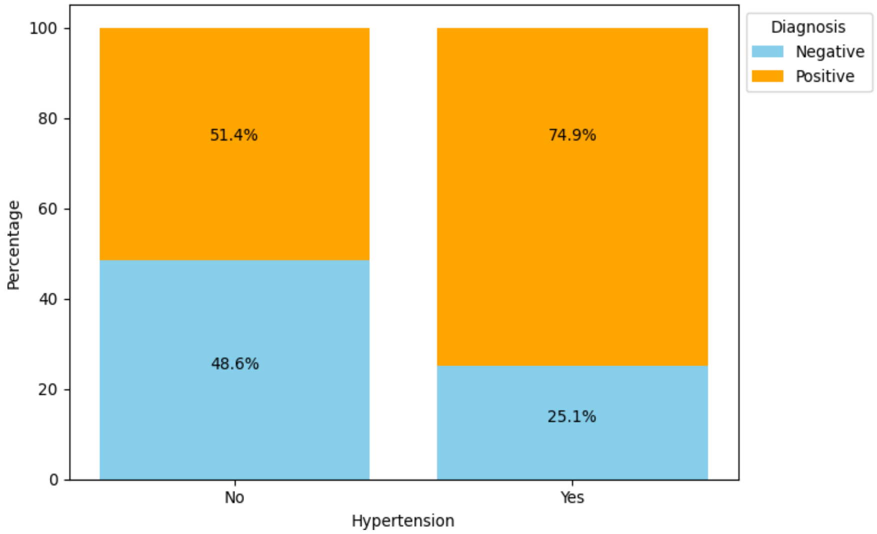

Figure 11: Distribution of Diagnosis by hypertension

The percentage of positive diagnoses increases significantly (from 51.4% to 74.9%) when hypertension is present. Individuals without hypertension have a nearly balanced distribution between positive and negative diagnoses, with a slight majority (51.4%).

Table 2: Information Gain on Categorical Features through Weight of Evidence

Feature	Total_IV	Interpretation
Gender	0.146839	Medium
Smoking	0.169154	Medium
GeneticRisk	0.175169	Medium
Diabetes	0.048520	Weak
Hypertension	0.127902	Medium

4.3 Correlation Analysis

This correlation matrix reveals that the variables in the liver cancer dataset have minimal correlation with one another. The correlation values are close to 0, indicating weak linear relationships. No strong multicollinearity is observed, meaning the variables are likely independent. Liver Function Test shows a weak positive correlation with BMI (0.0437) and weak negative correlations with other variables. Age, Alcohol Consumption, and Physical Activity show negligible correlations with one another or with Liver Function Test. This suggests the dataset variables may contribute independently to liver cancer prediction.

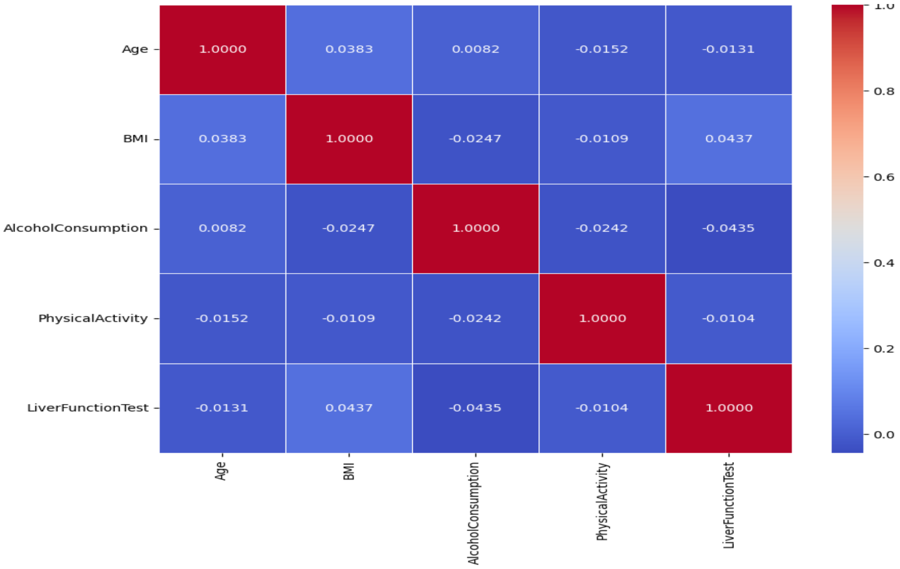

Figure 12: Correlation Matrix for The Features

5. CONCLUSIONS

This study provides valuable insights into the risk factors associated with liver cancer, emphasizing the role of age, BMI, alcohol consumption, smoking, genetic predisposition, diabetes, and hypertension. The findings indicate that middle-aged individuals, those with higher BMI, and individuals with higher alcohol consumption are at greater risk. Moreover, statistical tests confirmed significant dependencies between liver cancer diagnosis and multiple risk factors.

Utilizing Weight of Evidence (WOE) and Information Value (IV), this research highlights the key predictors, supporting in early detection and risk stratification. The correlation analysis revealed that while individual features exhibit minimal direct correlation, they independently contribute to liver cancer prediction.

These insights can support healthcare professionals in developing targeted screening programs and preventive strategies, ultimately improving early diagnosis and patient outcomes. Future research could extend these findings by incorporating Machine Learning models for more precise predictive analytics.

REFERENCES

1. Arbain, A. N., & Balakrishnan, B. Y. P. (2019, February 9). *A comparison of data mining algorithms for liver cancer prediction on imbalanced data*. International Journal of Data Science, Analytics and Applications. http://ijdsaa.com/index.php/welcome/article/view/2
2. Attiya, I. M., Abouelsoud, R. A., & Ismail, A. S. (2023, May 22). *A proposed approach for predicting liver cancer*. Natural Sciences Publishing. https://www.naturalspublishing.com/files/published/p2o85hg1229b2y.pdf

3. Baitharu, T. R., & Pani, S. K. (2016). *Analysis of data mining techniques for healthcare decision support*. Procedia Computer Science, 85, 535–542. https://www.sciencedirect.com/science/article/pii/S1877050916306263

4. David, M. J. (2023, January). *Liver cancer EDA*. Google Colab. https://colab.research.google.com/drive/14kA2j79EACKlk_7sfmBRho1g4Y4DLCh2

5. Ghosh, M., et al. (2021, May 2). *A comparative analysis of machine learning algorithms to predict liver cancer*. https://d1wqtxts1xzle7.cloudfront.net/68837787/TSP_IASC_44090.pdf

6. Ghosh, S. R., & Waheed, S. (2017). *Analysis of classification algorithms for liver cancer diagnosis*. Journal Binet. https://www.journalbinet.com/uploads/2/1/0/0/21005390/38_jstei_analysis_of_classification_algorithms_for_liver_cancer_diagnosis.pdf

7. Hemalatha, M., Naik, B. C., & Kullayappa, K. C. (2021, October). *Detection and comparative analysis of liver cancer using machine learning models*. https://www.researchgate.net/publication/361885679

8. Kefelegn, S., & Kamat, P. (2018). *Prediction and analysis of liver disorder cancers by using data mining technique: Survey*. https://www.researchgate.net/publication/323277681

9. Khaled, O. M., et.al., (2023, February). *Evaluating machine learning models for predictive analytics of liver cancer detection using healthcare big data*. https://portal.arid.my/Publications/57f0c8e0-5a8f-4a4c-b3e4-54b64234d623.pdf

10. Kumar, M. K., Sreedevi, M., & Reddy, Y. C. A. P. (2018). *Survey on machine learning algorithms for liver cancer diagnosis and prediction*. https://d1wqtxts1xzle7.cloudfront.net/115834528/3508-libre.pdf

11. Mostafa, F., Hasan, E., Williamson, M., & Khan, H. (2021, December 1). *Statistical machine learning approaches to liver*. Machine Learning and Knowledge Extraction, 1(4), 232–246. https://www.mdpi.com/2673-4389/1/4/23

12. Nabeel, M., Majeed, S., Awan, M. J., & Muslih-Ud-Din, H. (2021, September). *Review on effective cancer prediction through data mining techniques*. ResearchGate. https://www.researchgate.net/publication/355192182

13. Pakhale, H., & Xaxa, D. K. (2016, May–June). *A survey on diagnosis of liver cancer classification*. International Journal of Emerging Technologies. https://d1wqtxts1xzle7.cloudfront.net/47157247/IJET-V2I3P22.pdf

14. Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2011, May). *A critical study of selected classification algorithms for liver cancer diagnosis*. https://d1wqtxts1xzle7.cloudfront.net/108374083

15. Razali, N., Mustapha, A., Abd Wahab, M. H., Mostafa, S. A., & Rostam, S. K. (2018). *A data mining approach to prediction of liver cancers*. https://www.researchgate.net/publication/342268167

16. Shetty, P. J., & Satyanarayana, P. (2023, July 23). *Prediction performance of classification models for imbalanced liver cancer data*. https://www.researchgate.net/publication/377241034

17. Sultan, T. I., Khedr, A., & Sabry, S. (2012, July). *Biochemical markers of fibrosis for chronic liver cancer: Data mining-based approach*. https://d1wqtxts1xzle7.cloudfront.net/78796384

18. Vijayrani, S., & Dhayanand, S. (2015, April). *Liver cancer prediction using SVM and Naïve Bayes algorithms*. https://www.researchgate.net/publication/339551659