Classifying Familial Hypercholesterolaemia: A Tree-based Machine Learning Approach

Marshima Mohd Rosli[1], Jafhate Edward[2], Marcella Onn[3]
Yung-An Chua[4], Noor Alicezah Mohd Kasim[5], Hapizah Nawawi[6]
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA[1, 2, 3]
40450 UiTM, Shah Alam, Selangor, Malaysia[1, 2, 3]
Institute for Pathology, Laboratory and Forensic Medicine (I-PPerForM) [1, 4, 5, 6]
University Teknologi MARA, 47000 UiTM, Sungai Buloh, Selangor, Malaysia[1, 4, 5, 6]
Faculty of Medicine, University Teknologi MARA, 47000 UiTM, Sungai Buloh, Selangor, Malaysia[5, 6]

***Abstract*—Familial hypercholesterolaemia is the most common and serious form of inherited hyperlipidaemia. It has an autosomal dominant mode of inheritance, and is characterised by severely elevated low-density lipoprotein cholesterol levels. Familial hypercholesterolaemia is an important cause of premature coronary heart disease, but is potentially treatable. However, the majority of familial hypercholesterolaemia individuals are under-diagnosed and under-treated, resulting in lost opportunities for premature coronary heart disease prevention. This study aims to assess performance of machine learning algorithms for enhancing familial hypercholesterolaemia detection within the Malaysian population. We applied three machine learning algorithms (random forest, gradient boosting and decision tree) to classify familial hypercholesterolaemia among Malaysian patients and to identify relevant features from four well-known diagnostic instruments: Simon Broome, Dutch Lipid Clinic Criteria, US Make Early Diagnosis to Prevent Early Deaths and Japanese FH Management Criteria. The performance of these classifiers was compared using various measurements for accuracy, precision, sensitivity and specificity. Our results indicated that the decision tree classifier had the best performance, with an accuracy of 99.72%, followed by the gradient boosting and random forest classifiers, with accuracies of 99.54% and 99.52%, respectively. The three classifiers with Recursive Feature Elimination method selected six common features of familial hypercholesterolaemia diagnostic criteria (family history of coronary heart disease, low-density lipoprotein cholesterol levels, presence of tendon xanthomata and/or corneal arcus, family hypercholesterolaemia, and family history of familial hypercholesterolaemia) that generate the highest accuracy in predicting familial hypercholesterolaemia. We anticipate machine learning algorithms will enhance rapid diagnosis of familial hypercholesterolaemia by providing the tools to develop a virtual screening test for familial hypercholesterolaemia.**

Keywords*—Familial hypercholesterolaemia; predicting FH; machine learning algorithms; tree-based classifier

I. Introduction

Familial hypercholesterolaemia (FH) is the most common and serious form of inherited hyperlipidaemia, and is characterised by severely elevated low-density lipoprotein cholesterol (LDL-C) levels. It is an important cause of premature atherosclerosis and coronary heart disease (CHD), but is potentially treatable [1], [2]. Globally, the prevalence of heterozygous FH has been estimated at 1:200–1:500 [3]. However, the majority of FH individuals remain under-diagnosed and under-treated, resulting in lost opportunities for preventing premature CHD (pCHD).

In Malaysia, the prevalence of hypercholesterolaemia and severe hypercholesterolaemia is approximately 60% and 3%, respectively, and we have recently reported a high community prevalence of clinically diagnosed FH of 1:100 [4]. Further, FH was detected in about 35% of patients with pCHD [5]. With an estimated Malaysian population of 32 million, it is projected that at least 64,000–160,000 individuals are affected, the majority of whom are likely to be undiagnosed or inadequately treated. However, the prevalence of confirmed FH is not well established in Malaysia because DNA testing is costly and not commonly available in primary care clinics. Screening based on the lipid profile and LDL-C related measures is a reasonable alternative approach to assess the risk present, but contends with problems [6].

FH is usually diagnosed using four well-known diagnostic instruments: Simon Broome (SB;[7]), Dutch Lipid Clinic Criteria (DLCC; [8]), US Make Early Diagnosis to Prevent Early Deaths (US MEDPED;[9]) and Japanese FH Management Criteria (JFHMC; [10]). In Malaysia, the reports of FH are highly varied in terms of diagnostic method [11], due to lack of consensus in usage FH diagnostic criteria for screening of FH. Additionally, the input variables and the outcome of each diagnostic criteria are different, therefore, any attempt to combine multiple diagnostic criteria into one diagnostic criteria is not possible. According to the national standard guideline for management of dyslipidaemia, clinicians may use the DLCC, SB and US-MEDPED tools to diagnose patients [12]. A handful of Malaysian FH study groups [13], [14] already reported their research findings based on these diagnostic criteria [15], [16].

The above-mentioned FH diagnostic instruments are traditionally paper-based, and the diagnostic outcomes are manually scored by healthcare providers. This practice, however, has various well-known shortcomings that are typical of paper-based data collection systems, such as the expense of paper and space constraints for printing and storage. In addition, as diagnostic criteria specifically designed for Malaysians are still not available, the need to choose among multiple instruments of diagnostic criteria means that diagnosing FH has become time-consuming and laborious.

Machine learning techniques have been widely applied in the field of medical diagnostic applications because they can perform large-scale data analysis and predict a potential outcome efficiently [17], [18] [19], [20]. These techniques incorporate the use of artificial intelligence, which learns the dataset's patterns, and subsequently designs and trains a predictive model. The model seeks to make predictions on new data and is commonly used for classification, decision-making and rule-mining. Using these techniques can help predict and identify FH individuals who are at risk of developing pCHD, which in turn opens a major opportunity in healthcare.

Therefore, the goal of our research is to determine the most relevant features of the above-mentioned four diagnostic instruments that are useful in the diagnosis of FH in Malaysian patients, using machine learning models. We apply three classification models (random forest, gradient boosting and decision tree classifiers) with a recursive feature elimination (RFE) algorithm to perform feature selection by iteratively training a model, ranking features, and then removing the lowest ranking features. We anticipate that the pertinent features selected by the three classifiers will assist Malaysian FH study groups to construct a set of population-based diagnostic criteria for FH screening in upcoming studies.

The contributions of this paper are:

- We present a range of different tree-based machine learning approach with Recursive Feature Elimination method for detection of FH in Malaysian population.
- We use the largest number of primary health care records that contain a diagnosis of FH according to four well-known diagnostic instruments (DLCC, SB, JFHMC and US MEDPED) conducted in Malaysia.
- We determine the novel predictive features that are useful in the diagnosis of FH in Malaysian patients, using machine learning models.

II. Related Work

In this section, we start with related work that discuss studies on the prediction and classifications of FH using machine learning techniques. Then, we discuss recent studies that predict the presence of FH-causing genetic mutations. Finally, we discuss the importance of tree-based machine learning techniques that provides important insights to this research.

Several studies on the prediction and classifications of FH using machine learning techniques have been conducted by various researchers. For example, Shi et al. (2014) used logistic regression [21] to estimate the prevalence of FH and its treatment for adults in a random Chinese population and to assess the associated risk factors. They found that there was a high prevalence of phenotypic FH among those aged ⩾50 years, which suggests that FH is common and remains under-detected among Chinese population. Their findings were consistent with other researchers showing under-detection and under-treatment of FH in other countries [22], [23].

A group of researchers used random forest as a machine learning approach, with electronic health record data from Stanford Health Care and random forest classification for identification of potential FH patients [19], [24]. Their aims were to promote early diagnosis and timely intervention for high-risk pCHD patients with undiagnosed FH by using random forest for performing features of FH score.

Weng et al. (2015) used a stepwise logistic regression method [25] to predict FH, involving nine variables. The stepwise logistic regression was used to improve the identification of individuals in primary care settings who could be prioritised for further clinical assessment. The study also removed one of the variables, family history, which eventually resulted in significant improvement in discrimination.

Later, the same group of researchers published a new study of identifying and managing possible FH using SB criteria in primary care setting [26]. The study used six variables (demographic data, family medical history, physical signs, lipid characteristics and statin used in medication habits) and two methods (descriptive analysis and Wald's method). Their results showed 118 of 831 patients who were at least 18 years of age had blood total cholesterol levels >7.5 mmol/L, and 32 of them were without previous diagnosis of FH.

Pina et al. (2020) used three machine learning algorithms to predict the presence of FH-causing genetic mutations in two independent FH cohorts: a classification tree (CT), a gradient boosting machine (GBM) and a neural network (NN) [27]. They found that the three machine learning algorithms performed better than the clinical DLCC in predicting carriers of FH-causative mutations by evaluating the area under receiver operating curve (AUROC) parameter. This indicates that machine learning techniques may help the confirmation of FH, especially in the context of primary care or specialist clinics such as specialist lipid, cardiology or endocrinology clinics, which may prompt family cascade screening for detection of more FH among family members.

Although several techniques have been proposed to resolve the challenges associated with the prediction and classification of FH, we found that there is still a lack of research in predicting and classifying FH patients with machine learning techniques to determine important features of FH diagnostic criteria to diagnose FH. As mentioned earlier, only a few groups of researchers have apparently used random forest to predict FH, and none appear to have utilised other tree-based machine learning techniques such as decision tree and gradient boosting, which generally involve human-like algorithms that are compatible with all four diagnostic instruments. Moreover, there is a scarcity of reports on the use of different machine learning models in predicting FH in the local Malaysian population.

The tree-based machine learning techniques were widely used for solving classification problem in prediction of disease due to ability to deal with many clinical predictors of disease. Decision tree model is the most fundamental of the tree-based approach that able to generate human-understable rules without requiring much computational effort. Random forest model is an ensemble of decision trees, which utilise bagging aggregation approach to gain many trees and average over multiple trees for reducing the possibility of overfitting. Gradient boosting model is another variation of an ensemble method, which uses subsets of the original data to generate a series of average performing models and then "boosts" their performance by merging them using a specific cost function. Hence, the decision tree, gradient boosting and random forest models were explored in this study to detect FH. We expect that the outcome from the best classification model can be used to identify the relevant features that generate the highest accuracy in predicting FH, which potentially facilitate the development of Malaysian-based FH diagnostic criteria in future.

III. Materials and Methods

A. Study Design and Population

In this study, we used a secondary dataset containing 5248 individuals from all states in Malaysia, who were recruited from community health screening programmes and specialist lipid clinics in Malaysia, such as the Universiti Teknologi MARA (UiTM) Specialist Lipid Clinic, UiTM Cardiology Clinic and National Heart Institute (IJN), from 2011 to 2019. Individuals with secondary causes of hypercholesterolaemia, such as nephrotic syndrome, hypothyroidism, chronic kidney disease and cholelithiasis, were excluded from the study.

The dataset consists of 24 raw features, with 54.05% of the dataset having complete fields. Because of the low percentage of complete fields, we applied univariate and multiple imputation methods to replace the quantitative missing values to overcome the limitation of missing values in the dataset. After the missing data were successfully imputed, the dataset was further processed to reduce the number of features with weak relations with the target feature.

The cleaned dataset comprised 16 features describing the patients' demographic data and clinical characteristics: age; gender; smoking habit; patient history of pCHD, cerebrovascular accident (CVA) or peripheral vascular disease (PVD), and diabetes; lipid profile including high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglycerides (TG); family history of FH, hypercholesterolaemia and pCHD; patient's physical symptoms of corneal arcus and tendon xanthomata; and whether the patient was on lipid-lowering therapy. Table I shows the demographic and clinical characteristics of the study population.

B. Decision Tree Approach and Algorithm

The experiments were conducted using SPSS Modeler 18 and Python. Three classification models were used to train and test the dataset: random forest, gradient boost and decision tree. The cleaned dataset was partitioned into 70:30 ratios for training and testing; 70% (3674 instances) of the overall dataset were labelled X_train and used to train the classification model, and 30% (1574 instances) of the dataset were labelled X_test and used to test the model.

TABLE I. Demographic and Clinical Characteristics of the Study Population (n = 5248)

Feature	Total
Age, mean (SD)	41.41±15.404
Gender	
Male	(2009) 38.3%
Female	(3238) 61.7%
High-density lipoprotein cholesterol, mean (SD)	1.29±0.40
Baseline low-density lipoprotein cholesterol, mean (SD)	3.27±1.14
Triglycerides cholesterol, mean (SD)	1.69±1.17
Total cholesterol, mean (SD)	5.32±1.43
Tendon xanthomata	(22) 0.4%
Corneal arcus	(263) 5.0%
Lipid-lowering therapy	(383) 7.3%
Smoking	(630) 12.0%
Diabetes	(342) 6.5%
History of coronary heart disease	(104) 2.0%
History of cerebrovascular accident or peripheral vascular disease	(64) 1.2%
Family history of familial hypercholesterolaemia	(84) 1.6%
Family history of hypercholesterolaemia	(728) 13.9%
Family history of coronary heart disease	(682) 13.0%

In this study, we used multi-class classification for the DLCC and SB because these diagnostic instruments involve classifying into one of more than two classes. We used binary classification for the JFHMC and US MEDPED diagnostic criteria because these diagnostic instruments classify into one of two classes. We applied an RFE algorithm with the three classification models to select a subset of the most relevant features for the dataset and to eliminate weak features identified as noises, which might affect the performance of the models. The RFE approach consisted of three steps: (a) training the classification model to determine initial importance scores, (b) removing the bottom features with the lowest importance scores from the dataset, and (c) assigning ranks to remove features according to the sequence of their most recent importance scores. These steps were executed iteratively until the specified number of remaining features rounded to zero.

We evaluated the performance of each classification model according to accuracy, sensitivity, specificity and precision values. The accuracy values were calculated using Eq. (1).

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP+TN+FN)} \quad (1)$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative based on a confusion matrix. We used sensitivity, specificity and precision values to support the accuracy values. The sensitivity and specificity methods are described in Eq. (2) and Eq. (3), respectively.

$$\text{Sensitivity } = \frac{TP}{(TP+FP)} \quad (2)$$

where TP is true positive, TN is true negative and FP is false positive based on a confusion matrix.

$$\text{Specificity } = \frac{TN}{(TN+FP)} \quad (3)$$

where TN is true negative and FP is false positive based on a confusion matrix.

$$\text{Precision } = \frac{TP}{(TP+FP)} \quad (4)$$

where TP is true positive and FP is false positive based on a confusion matrix.

IV. Results

The best model was based on the highest accuracy value supported by sensitivity, specificity and precision values. Table II shows the results for predictive accuracy values for each diagnostic instrument and model.

Overall, results show that all the models can be used for classifying the FH dataset. However, in overall performance, the decision tree model produced the highest accuracy value. The model recorded an impressive average accuracy value of 99.72% compared with the other models (random forest, 99.54%, and gradient boosting, 99.52%). Of note, the accuracy values obtained by the decision tree model for the DLCC and JFHMC diagnostic instruments were the main contributors to its overall performance. The model obtained perfect accuracy values of 100% for the DLCC diagnostic instrument, in which it outperformed the other models because of some advantages of the decision tree model, such as splitting criteria and the pruning method. The multi-way splitting tree of the decision tree model was advantageous when dealing with multi-classification involving more than two classes.

Table III shows the results for sensitivity, specificity and precision values for each model across the four diagnostic instruments. Overall, results show that all the models can be used to classify FH patients correctly according to the DLCC, US MEDPED and JFHMC diagnostic tools. However, for SB diagnostic criteria, the models encountered a problem caused by two factors: (1) multi-classification involving three classes and (2) high similarity of data.

According to the sensitivity results for the DLCC, the decision tree model demonstrated the perfect value (100%). The gradient boosting model was fairly close, with a value of 75%, while the random forest model was rated 43.75%. For the JFHMC, all the models demonstrated the perfect value (100%) for sensitivity. For the US MEDPED, the random forest model achieved the highest sensitivity value with 99.81% compared with the gradient boosting model (99.48%) and random forest model (99.55%).

TABLE II. Classification of Accuracy Values for Machine Learning Models Across the Four Diagnostic Instruments

Accuracy (%)			
Diagnostic instrument	Decision tree	Random forest	Gradient boosting
DLCC	100.00	99.36	99.49
SB	99.75	99.81	99.74
JFHMC	100.00	99.94	100.00
US MEDPED	99.11	99.05	98.86
Average	99.72	99.54	99.52

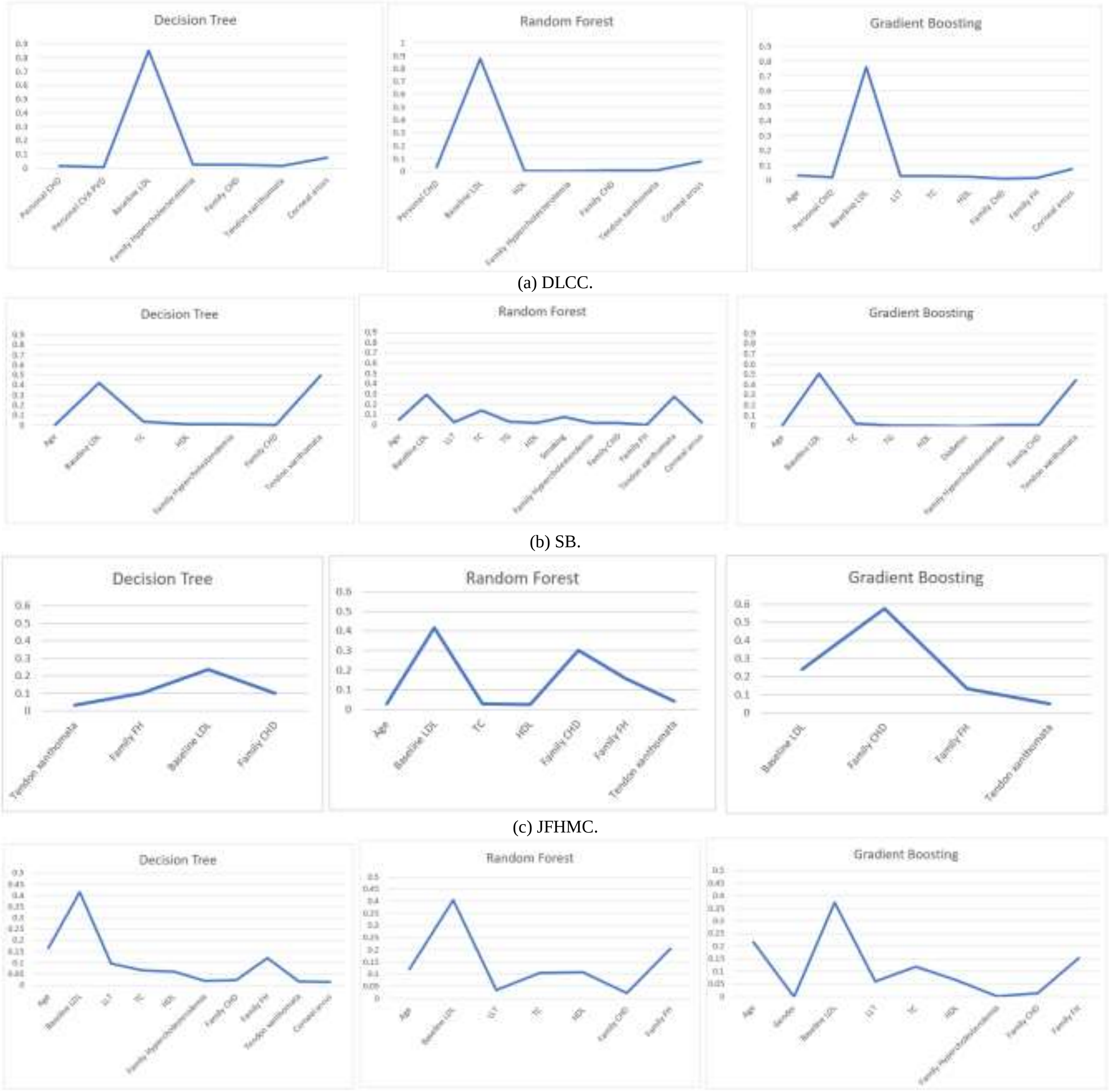
SB: Simon Broome diagnostic criteria; DLCC: Dutch Lipid Clinic Criteria; JFHMC: Japanese FH Management Criteria; US MEDPED: US Make Early Diagnosis to Prevent Early Deaths.

TABLE III. Classification of Sensitivity, Specificity and Precision Values for Machine Learning Models Across four well-known Diagnostic Instruments

Machine learning model	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**No. of features**
DLCC					
Random forest	99.36%	43.75%	99.94%	87.50%	7
Gradient boosting	99.49%	75.00%	99.74%	75.00%	9
Decision tree	**100%**	**100%**	**100%**	**100%**	**7**
SB					
Random forest	99.81%	25.00%	100%	100%	12
Gradient boosting	99.74%	100%	99.74%	50.00%	9
Decision tree	99.75%	0%	100%	0%	7
JFHMC					
Random forest	99.94%	100%	98.63%	99.93%	7
Gradient boosting	100%	100%	100%	100%	4
Decision tree	**100%**	**100%**	**100%**	**100%**	**4**
US MEDPED					
Random forest	99.05%	99.81%	65.71%	99.23%	7
Gradient boosting	98.86%	99.48%	71.43%	99.35%	9
Decision tree	**99.11%**	**99.55%**	**80.00%**	**99.55%**	**10**

Specificity values ranged from 65% for random forest to 80% for decision tree for the US MEDPED. Decision tree had perfect specificity values (100%) for the DLCC, SB and JFHMC, and gradient boosting had perfect specificity for the JFHMC. A precision value of 100% was obtained by decision tree for the DLCC and JFHMC, random forest for the SB, and gradient boosting for the JFHMC.

Based on accuracy, sensitivity, specificity and precision values, decision tree is the best model for classifying the FH dataset according to the diagnostic criteria of the DLCC, JFHMC and US MEDPED instruments. For further verification, Fig. 1 shows the clinical feature ranking by feature importance using RFE for the four diagnostic tools (DLCC, SB, JFHMC and US MEDPED) across the three classification models. Each classification model was run on RFE, which was initiated with one clinical feature and increased the number of clinical features until it reached the maximum number. The best model was mainly based on the highest accuracy value and the minimum number of clinical features for the specific tree-based model.



(a) DLCC.

(b) SB.

(c) JFHMC.

(d) US MEDPED.

Fig. 1. Feature Ranking by Feature Importance for Four Diagnostic Instruments across Three Classification Models.

In Fig. 1(a), the DLCC shows perfect accuracy (100%) reached by the decision tree classifier with seven clinical features (history of CHD, history of CVA or PVD, family history of hypercholesterolaemia, family history of CHD, presence of tendon xanthomata, and presence of corneal arcus, LDL-C level) included in the model. Fig. 1(b) shows that the maximum accuracy for the SB reached by the random forest classifier is 99.81%, with 12 clinical features (age, total cholesterol, triglycerides cholesterol level, LDL-C level, lipid-lowering therapy, smoking habit, HDL-C level, family history of hypercholesterolaemia, family history of CHD, family history of FH, presence of tendon xanthomata, and presence of corneal arcus) included in the model.

In Fig. 1(c), the JFHMC shows the maximum accuracy reached by the decision tree classifier and gradient boosting is 100%, with four clinical features (family history of CHD, family history of FH, LDL-C level, presence of tendon xanthomata) included in the model. In Fig. 1(d), the US MEDPED shows the maximum accuracy reached by the random forest classifier is 99.11%, with seven features (age, lipid-lowering therapy, total cholesterol, HDL-C level, family history of CHD, family history of FH, LDL-C level) included in the model. Overall results show that the decision tree classifier (Fig. 1) outperformed the other classifiers in terms of accuracy and minimum numbers of clinical features selected.

In terms of number of features in the dataset, the random forest classifier showed the most selected features (12) for SB criteria compared with the other models, which mostly had seven features selected across the four diagnostic instruments. An increase in features is indicative of a longer period taken for the model to process. Therefore, fewer features are preferable for significant improvement of accuracy performance, comprising the strongest features identified by RFE. The three classification models with RFE selected six common features: family history of CHD, LDL-C level, presence of tendon xanthomata and/or corneal arcus, family hypercholesterolaemia, and family history of FH. Overall, we found that the decision tree classifier is the best model for classification as it demonstrated the highest accuracy and selected the minimum number of features among the classification models. Based on our results, the best diagnostic instrument is the one that includes the maximum number of the six relevant features that can help to accurately classify FH patients. Table IV shows the presence or absence of the selected six features across each diagnostic instrument.

TABLE IV. PRESENCE OF SELECTED FEATURES IN EACH DIAGNOSTIC INSTRUMENT

Selected features	SB	DLCC	US MEDPED	JFHMC
Family history of CHD	√	√	X	√
LDL-C level	√	√	X	√
Family history of hypercholesterolaemia	X	√	X	X
Family history of FH	X	X	√	√
Presence of tendon xanthomata	√	√	X	√
Presence of corneal arcus	X	√	X	X

From Table IV, the DLCC instrument includes five of the six selected features (except family history of FH), and the JFHMC instrument includes four of the features (except corneal arcus and family history of hypercholesterolaemia). None of the features were present in the US MEDPED criteria, except family history of FH. This indicates that the DLCC instrument is the most suitable for Malaysian patients on the basis of the relevant features selected by the best classification model.

V. DISCUSSION

This study is the first to report on detection of FH by applying machine learning models (random forest, gradient boosting and decision tree) with RFE to over 5000 primary health care records that contain a clinically diagnosis of FH according to four well-known diagnostic instruments (DLCC, SB, JFHMC and US MEDPED). Machine learning models provide an additional effective way of screening patients and do not replace the clinical evaluation using diagnostic criteria.

In our study, results showed that the three machine learning models had similar high predictive accuracy in classifying FH patients (accuracy > 99.00%). This is consistent with prior findings using a random forest algorithm in health data [19] and other prior findings using random forest, gradient boosting, deep learning and ensemble learning algorithms in primary care data [28]. The decision tree model outperformed the other machine learning models, with the highest accuracy to determine the likelihood of FH.

Despite the similar accuracy, this study found minimal differences for other performance values between machine learning models. Our analysis highlights specificity values were consistently high across all machine learning models for DLCC, SB and JFHMC that indicate the proportion of patients without actual FH were correctly classified. However, results for sensitivity and precision values varied between machine learning models. For example, random forest model for DLCC identified small proportion of patients with actual FH due to the low sensitivity value (43.75%), but the model would be efficient in having a higher detection rate of FH (high precision value 87.5%).

This study further highlights variations in the selected clinical features identified by the different machine learning models used. For example, decision tree for the DLCC identified seven clinical features (history of CHD, history of CVA or PVD, LDL-C level, family history of hypercholesterolaemia, family history of CHD, presence of tendon xanthomata, presence of and corneal arcus), which is in line with the SB and DLCC diagnostic criteria to systematically identify those who are likely to have FH. Gradient boosting and decision tree for the JFHMC identified four clinical features (family history of CHD, family history of FH, LDL-C level and presence of tendon xanthomata), and random forest for the US MEDPED identified seven features (age, lipid-lowering therapy, total cholesterol, HDL-C level, family history of CHD, family history of FH and LDL-C level). Taken together, these results suggest six relevant clinical features across four diagnostic instruments that can predict FH in Malaysian population: family history of CHD, LDL-C level, presence of tendon xanthomata, presence of corneal arcus, family history of hypercholesterolaemia and family history of FH.

The findings of this study have important implications for developing FH diagnostic criterion specific for Malaysian population. Our study suggest that machine learning models allow the identification of novel predictive features for detecting FH in Malaysian population. For instance, five out of the six relevant features are well-established criteria in the DLCC diagnostic instrument [8] which previous studies on FH in Malaysia applied DLCC as the main reference diagnostic criteria and it is widely recommended globally [4], [15]. Future studies, which take these novel predictive features into account, will be undertaken.

This study recommends several strengths. We evaluate a range of different tree-based machine learning approach with Recursive Feature Elimination method for detection of FH in Malaysian population. We used the largest number of primary health care records that contain a diagnosis of FH according to four well-known diagnostic instruments (DLCC, SB, JFHMC and US MEDPED), compared to other previous FH studies conducted in Malaysia. This study also assessed the clinical features of the abovementioned four diagnostic instruments to identify the novel predictive features that are useful in the diagnosis of FH in Malaysian patients, using machine learning models.

Compared with relying on multiple FH diagnostic criteria, as being practised currently, the use of machine learning techniques allows healthcare providers to conduct early testing for the presence of FH in patients. It simplifies the current labour-intensive and time-consuming process in the diagnosis of FH in Malaysia by streamlining and focusing on important features of diagnostic criteria that are relevant and pertinent to the procedure. The machine learning techniques offer major opportunities to increase diagnosis of FH and to prevent pCHD and early death.

However, we acknowledge several study limitations, which are common in other research using healthcare data. The limitations include the potential for information bias due to missing data. Missing data may introduce bias in the performance of prediction models. However, we used mean or mode imputation methods to replace quantitative missing values with the mean of the attribute or qualitative missing values with the mode of the attribute to overcome these effects. Another potential information bias in the dataset is that some patients could potentially be misclassified because of inaccurate reporting of family history. Future studies should validate and replicate our machine learning models with the implementation of RFE in other populations to confirm the findings of this study. Further, additional evaluation of the feasibility of machine learning applications in clinical practice is required to support the computational capacity of healthcare systems.

VI. Conclusion

The decision tree classifier performs best in identifying the relevant features for the DLCC, SB, US MEDPED and JFHMC. Family history of CHD, family history of hypercholesterolemia, family history of FH, LDL-C level, presence of tendon xanthomata and presence of corneal arcus, are the relevant features for diagnosing FH among DLCC, SB, US MEDPED and JFHMC diagnostic criteria that give the highest accuracy in the classification model. Future research should include these six relevant features, which have potential to be developed into an efficient FH prediction model to assist clinicians in identifying FH patients.

Overall, this study highly suggests that machine learning algorithms may help the diagnosis of FH in classifying FH among patients, leading to effective identification of high-risk patients with FH. The three classifiers used in this study embody the most important features in predicting patients with FH. These features also contribute to unify the population-based diagnostic criteria, constituting a first step towards development of more relevant, locally adjusted and tested Malaysian FH diagnostic criteria for early diagnosis of FH in the local community. This is also particularly important in family contact tracing for indexed cases. Efficient, locally adjusted diagnostic criteria will improve early and overall detection, hence anticipating early treatment and prevention of pCHD.

Acknowledgment

The authors would like to thank the Universiti Teknologi Mara for their financial support of this project under FRGS Grant No. 600-IRMI/FRGS 5/3 (212/2019).

References

[1] T. Phuong Kim, L. Thuan Duc, and H. Le Thuy Ai, “The Major Molecular Causes of Familial Hypercholesterolemia,” Asian J. Pharm. Res. Heal. Care, vol. 10, no. 2, pp. 60–68, Aug. 2018, doi: 10.18311/ajprhc/2018/20031.

[2] A. Wiegman, S. S. Gidding, G. F. Watts, M. J. Chapman, H. N. Ginsberg, M. Cuchel, L. Ose, M. Averna, C. Boileau, J. Borén, E. Bruckert, A. L. Catapano, J. C. Defesche, O. S. Descamps, R. A. Hegele, G. K. Hovingh, S. E. Humphries, P. T. Kovanen, J. A. Kuivenhoven, L. Masana, B. G. Nordestgaard, P. Pajukanta, K. G. Parhofer, F. J. Raal, K. K. Ray, R. D. Santos, A. F. H. Stalenhoef, E. Steinhagen- Thiessen, E. S. Stroes, M.-R. Taskinen, A. Tybjærg-Hansen, and O. Wiklund, “Familial hypercholesterolaemia in children and adolescents: gaining decades of life by optimizing detection and treatment,” Eur. Heart J., vol. 36, no. 36, pp. 2425–2437, Sep. 2015, doi: 10.1093/eurheartj/ehv157.

[3] B. G. Nordestgaard, M. J. Chapman, S. E. Humphries, H. N. Ginsberg, L. Masana, O. S. Descamps, O. Wiklund, R. A. Hegele, F. J. Raal, J. C. Defesche, A. Wiegman, R. D. Santos, G. F. Watts, K. G. Parhofer, G. K. Hovingh, P. T. Kovanen, C. Boileau, M. Averna, J. Borén, E. Bruckert, A. L. Catapano, J. A. Kuivenhoven, P. Pajukanta, K. Ray, A. F. H. Stalenhoef, E. Stroes, M.-R. Taskinen, A. Tybjærg-Hansen, and European Atherosclerosis Society Consensus Panel, “Familial hypercholesterolaemia is underdiagnosed and undertreated in the general population: guidance for clinicians to prevent coronary heart disease: consensus statement of the European Atherosclerosis Society.,” Eur. Heart J., 2013, doi: 10.1093/eurheartj/eht273.

[4] Y.-A. Chua, A. Z. Razman, A. S. Ramli, N. A. Mohd Kasim, and H. M. Nawawi, “Familial Hypercholesterolaemia in the Malaysian Community: Prevalence, Under-Detection and Under-Treatment,” J. Atheroscler. Thromb., 2021, doi: 10.5551/jat.57026.

[5] S. A. Nazli, Y. A. Chua, N. A. Mohd Kasim, Z. Ismail, A. B. Md Radzi, K. S. Ibrahim, S. Kasim, A. Rosman, and H. M. Nawawi, “Familial Hypercholesterolaemia among Patients with Coronary Angiogram-Proven Premature Coronary Artery Disease,” Strait Circ. J., vol. 1, no. 2, p. 44, 2019, doi: https://doi.org/10.6907/SCJ.201909/SP_1(2).0037.

[6] J. C. Defesche, “Defining the challenges of FH Screening for familial hypercholesterolemia,” J. Clin. Lipidol., vol. 4, no. 5, pp. 338–341, Sep. 2010, doi: 10.1016/j.jacl.2010.08.022.

[7] K. E. Heath, S. E. Humphries, H. Middleton-Price, and M. Boxer, “A molecular genetic service for diagnosing individuals with familial hypercholesterolaemia (FH) in the United Kingdom,” Eur. J. Hum. Genet., vol. 9, no. 4, pp. 244–252, Apr. 2001, doi: 10.1038/sj.ejhg.5200633.

[8] S. W. Fouchier, J. C. Defesche, M. A. Umans-Eckenhausen, and J. J. Kastelein, “The molecular basis of familial hypercholesterolemia in The Netherlands,” Hum. Genet., vol. 109, no. 6, pp. 602–615, Dec. 2001, doi: 10.1007/s00439-001-0628-8.

[9] R. R. Williams, S. C. Hunt, M. C. Schumacher, R. A. Hegele, M. F. Leppert, E. H. Ludwig, and P. N. Hopkins, “Diagnosing heterozygous familial hypercholesterolemia using new practical criteria validated by molecular genetics,” Am. J. Cardiol., vol. 72, no. 2, pp. 171–176, Jul. 1993, doi: 10.1016/0002-9149(93)90155-6.

[10] M. Harada-Shiba, H. Arai, S. Oikawa, T. Ohta, T. Okada, T. Okamura, A. Nohara, H. Bujo, K. Yokote, A. Wakatsuki, S. Ishibashi, and S. Yamashita, “Guidelines for the management of familial hypercholesterolemia,” J. Atheroscler. Thromb., 2012, doi: 10.5551/jat.14621.

[11] A. Al-Khateeb and H. Al-Talib, “Genetic Researches Among Malaysian Familial Hypercholesterolaemia Population,” J. Heal. Transl. Med., vol. 19, no. 2, pp. 1–11, Dec. 2016, doi: 10.22452/jummec.vol19no2.1.

[12] R. Jeyamalar, W. A. Wan Azman, H. Nawawi, G. H. Choo, W. K. Ng, M. A. Rosli, O. Al Fazir, K. Sazzli, M. Oteh, and D. K. L. Quek, “Updates in the management of Dyslipidaemia in the high and very high risk individual for CV risk reduction.,” Med. J. Malaysia, vol. 73, no. 3, pp. 154–162, 2018, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29962499.

[13] K. L. Khoo, P. Van Acker, H. Tan, and J. P. Deslypere, “Genetic causes of familial hypercholesterolaemia in a Malaysian population.,” Med. J. Malaysia, vol. 55, no. 4, pp. 409–18, Dec. 2000, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11221151.

[14] S.-H. Lye, J. K. Chahil, P. Bagali, L. Alex, J. Vadivelu, W. A. W. Ahmad, S.-P. Chan, M.-K. Thong, S. M. Zain, and R. Mohamed, “Genetic Polymorphisms in LDLR, APOB, PCSK9 and Other Lipid Related Genes Associated with Familial Hypercholesterolemia in Malaysia,” PLoS One, vol. 8, no. 4, p. e60729, Apr. 2013, doi: 10.1371/journal.pone.0060729.

[15] S. Abdul-Razak, R. Rahmat, A. Mohd Kasim, T. A. Rahman, S. Muid, N. M. Nasir, Z. Ibrahim, S. Kasim, Z. Ismail, R. Abdul Ghani, A. R. Sanusi, A. Rosman, and H. Nawawi, “Diagnostic performance of various familial hypercholesterolaemia diagnostic criteria compared to Dutch lipid clinic criteria in an Asian population,” BMC Cardiovasc. Disord., vol. 17, no. 1, p. 264, Dec. 2017, doi: 10.1186/s12872-017-0694-z.

[16] A. Al-Khateeb, M. K. Zahri, M. S. Mohamed, T. H. Sasongko, S. Ibrahim, Z. Yusof, and B. A. Zilfalil, “Analysis of sequence variations in low-density lipoprotein receptor gene among Malaysian patients with familial hypercholesterolemia,” BMC Med. Genet., vol. 12, no. 1, p. 40, Dec. 2011, doi: 10.1186/1471-2350-12-40.

[17] A. Khan, J. P. Li, A. U. Haq, I. Memon, S. H. Patel, and S. ud Din, “Emotional-physic analysis using multi-feature hybrid classification,” J. Intell. Fuzzy Syst., vol. 40, no. 1, 2021, doi: 10.3233/JIFS-201069.

[18] M. H. Memon, I. Memon, J. P. Li, and Q. A. Arain, “IMRBS: image matching for location determination through a region-based similarity technique for CBIR*,” Int. J. Comput. Appl., vol. 41, no. 6, 2019, doi: 10.1080/1206212X.2018.1468643.

[19] K. D. Myers, J. W. Knowles, D. Staszak, M. D. Shapiro, W. Howard, M. Yadava, D. Zuzick, L. Williamson, N. H. Shah, J. M. Banda, J. Leader, W. C. Cromwell, E. Trautman, M. F. Murray, S. J. Baum, S. Myers, S. S. Gidding, K. Wilemon, and D. J. Rader, “Precision screening for familial hypercholesterolaemia: a machine learning study applied to electronic health encounter data,” Lancet Digit. Heal., vol. 1, no. 8, pp. e393–e402, Dec. 2019, doi: 10.1016/S2589-7500(19)30150-5.

[20] Z. Obermeyer and E. J. Emanuel, “Predicting the Future — Big Data, Machine Learning, and Clinical Medicine,” N. Engl. J. Med., vol. 375, no. 13, pp. 1216–1219, Sep. 2016, doi: 10.1056/NEJMp1606181.

[21] Z. Shi, B. Yuan, D. Zhao, A. W. Taylor, J. Lin, and G. F. Watts, “Familial hypercholesterolemia in China: Prevalence and evidence of underdetection and undertreatment in a community population,” Int. J. Cardiol., 2014, doi: 10.1016/j.ijcard.2014.04.165.

[22] M. Benn, G. F. Watts, A. Tybjaerg-Hansen, and B. G. Nordestgaard, “Familial hypercholesterolemia in the Danish general population: Prevalence, coronary artery disease, and cholesterol-lowering medication,” J. Clin. Endocrinol. Metab., 2012, doi: 10.1210/jc.2012-1563.

[23] G. F. Watts, S. Gidding, A. S. Wierzbicki, P. P. Toth, R. Alonso, W. V. Brown, E. Bruckert, J. Defesche, K. K. Lin, M. Livingston, P. Mata, K. G. Parhofer, F. J. Raal, R. D. Santos, E. J. G. Sijbrands, W. G. Simpson, D. R. Sullivan, A. V. Susekov, B. Tomlinson, A. Wiegman, S. Yamashita, and J. J. P. Kastelein, “Integrated guidance on the care of familial hypercholesterolaemia from the International FH Foundation.,” Int. J. Cardiol., vol. 171, no. 3, pp. 309–25, Feb. 2014, doi: 10.1016/j.ijcard.2013.11.025.

[24] J. M. Banda, A. Sarraju, F. Abbasi, J. Parizo, M. Pariani, H. Ison, E. Briskin, H. Wand, S. Dubois, K. Jung, S. A. Myers, D. J. Rader, J. B. Leader, M. F. Murray, K. D. Myers, K. Wilemon, N. H. Shah, and J. W. Knowles, “Finding missed cases of familial hypercholesterolemia in health systems using machine learning,” npj Digit. Med., 2019, doi: 10.1038/s41746-019-0101-5.

[25] S. F. Weng, J. Kai, H. Andrew Neil, S. E. Humphries, and N. Qureshi, “Improving identification of familial hypercholesterolaemia in primary care: Derivation and validation of the familial hypercholesterolaemia case ascertainment tool (FAMCAT),” Atherosclerosis, vol. 238, no. 2, pp. 336–343, Feb. 2015, doi: 10.1016/j.atherosclerosis.2014.12.034.

[26] S. Weng, J. Kai, J. Tranter, J. Leonardi-Bee, and N. Qureshi, “Improving identification and management of familial hypercholesterolaemia in primary care: Pre- and post-intervention study,” Atherosclerosis, vol. 274, pp. 54–60, Jul. 2018, doi: 10.1016/j.atherosclerosis.2018.04.037.

[27] A. Pina, S. Helgadottir, R. M. Mancina, C. Pavanello, C. Pirazzi, T. Montalcini, R. Henriques, L. Calabresi, O. Wiklund, M. P. Macedo, L. Valenti, G. Volpe, and S. Romeo, “Virtual genetic diagnosis for familial hypercholesterolemia powered by machine learning,” Eur. J. Prev. Cardiol., p. 204748731989895, 2020, doi: 10.1177/2047487319898951.

[28] R. K. Akyea, N. Qureshi, J. Kai, and S. F. Weng, “Performance and clinical utility of supervised machine-learning approaches in detecting familial hypercholesterolaemia in primary care,” npj Digit. Med., vol. 3, no. 1, p. 142, Dec. 2020, doi: 10.1038/s41746-020-00349-5.