MINISTRY OF EDUCATION AND TRAINING MINISTRY OF HEALTH

HANOI MEDICAL UNIVERSITY

DO DUC HUY

APPLYING MACHINE LEARNING MODELS IN PRENATAL SCREENING FOR DOWN SYNDROME AT HANOI MEDICAL UNIVERSITY HOSPITAL

MASTER THESIS IN EPIDEMIOLOGY

HANOI - 2023

MINISTRY OF EDUCATION AND TRAINING MINISTRY OF HEALTH

HANOI MEDICAL UNIVERSITY

DO DUC HUY

APPLYING MACHINE LEARNING MODELS IN PRENATAL SCREENING FOR DOWN SYNDROME AT HANOI MEDICAL UNIVERSITY HOSPITAL

Major: Epidemiology

Code: 8720117

MASTER THESIS

Supervisor

Assoc.Prof. Le Minh Giang

Assoc.Prof. Nguyen Thi Trang

HANOI - 2023

Contents

Acknowledgement

Commitment

Abbreviation

List of Tables

List of Figures

Introduction **1**

1 Chapter 1: Literature Review **4**

1.1 **Down Syndrome** 4

1.1.1 **Overview** 4

1.1.2 **Genetic Features of Down Syndrome** 4

1.1.3 **Prenatal screening methods** 5

1.1.3.1 **Ultrasound** 5

1.1.3.2 **Double test, Triple test** 5

1.1.3.3 **Non-invasive prenatal screening (NIPT)** 8

1.1.4 **Diagnosing Down Syndrome** 8

1.1.4.1 **Amniocentesis (golden standard of this study)** 8

1.1.4.2 **Chorionic Villus Sampling** 9

1.2 **Artificial Intelligence (AI)** 9

1.2.1 **Definition** 9

1.2.2 **Machine learning** 9

1.2.2.1 **Supervised learning** 10

1.2.2.2 **Unsupervised learning** 11

1.2.3 **Machine learning algorithms** 12

1.2.3.1 **K nearest neighbor (kNN)** 12

1.2.3.2 **Random forest (RF)** . . . 12
1.2.3.3 **Support vector machine (SVM)** . . . 13
1.2.3.4 **Extreme Gradient Boosting (XGBoost)** . . . 13
1.2.4 **AI application in health care** . . . 14
1.2.5 **Studies applying AI in prenatal screening for Down syndrome** 15
1.3 **Evaluating the effectiveness of artificial intelligence software** . . . 17
1.3.1 **Dataset and model evaluation** . . . 17
1.3.2 **Evaluation parameters** . . . 18
1.3.2.1 **Accuracy** . . . 18
1.3.2.2 **Confusion Matrix (2x2 table)** . . . 19
1.3.2.3 **ROC curve** . . . 20

2 Chapter 2: Study Subjects And Research Methodology 23
2.1 **Description of the parent study** . . . 23
2.1.1 **Study timeline** . . . 23
2.1.2 **Study participants** . . . 23
2.1.3 **Sample size and sampling methods** . . . 24
2.1.4 **Variables used to build machine learning models** . . . 24
2.1.5 **Data collection tools used** . . . 24
2.1.6 **Potential source of biases and ways to prevent their occurrence** 25
2.1.6.1 **Source of biases** . . . 25
2.1.6.2 **Ways to prevent biases** . . . 25
2.1.7 **Ethical issues** . . . 25
2.2 **Description of the present study** . . . 25
2.2.1 **Study diagram** . . . 26
2.2.2 **Extracted information** . . . 26
2.2.3 **Statistical analysis** . . . 28
2.2.3.1 **Assessing sensitivity, specificity of AI models** . . . 28

3 Chapter 3: Results **31**

3.1 **The process of making machine learning models** 31

3.1.1 **Data overview** 31

3.1.2 **Characteristics of participants in each dataset** 34

3.1.3 **Features of each dataset** 36

3.1.4 **Model performance of each machine learning model in each dataset** 42

3.1.4.1 **Trimester 1** 42

3.1.4.2 **Trimester 2** 44

3.2 **Performance of each machine learning model on the test set** 46

3.2.1 **Trimester 1** 46

3.2.2 **Trimester 2** 48

4 Chapter 4: Discussion **50**

4.1 **Dataset characteristics** 50

4.1.1 **Included variables** 50

4.1.2 **Training, validating and testing dataset** 51

4.1.3 **Study population characteristics** 53

4.2 **Feature importance** 54

4.2.1 **Trimester 1** 54

4.2.2 **Trimester 2** 55

4.3 **Model performance** 56

4.3.1 **Hyper-parameters used** 56

4.3.2 **Model accuracy** 58

4.3.2.1 **Better result by adding more variables** 58

4.3.2.2 **Better result by pre-processing data** 58

4.3.2.3 **The problem with small number of test cases** 60

4.3.2.4 **Comparison with current Down screening method** . . . 60

4.4 **Applications** 61

4.5 **Study strengths and limitations** 63
4.5.1 **Study strengths** 63
4.5.2 **Study limitations** 63

5 Chapter 4: Conclusions **65**

6 Chapter 5: Recommendations **66**

References **67**

Acknowledgement

I would like to offer my thanks to the staffs of Hanoi Medical University, Institute of Preventive Medicine and Public Health, as well as teachers from the Department of Epidemiology for your guidance and support.

I would like to express my great appreciation to all the individuals and teams below, that without them, I would not be able to accomplish my Master thesis.

Firstly, I would like to express my deepest gratitude to my supervisors, Assoc.Prof Le Minh Giang and Assoc.Prof Nguyen Thi Trang for their guidance. Even though they has always been busy, they still willing to give me some of his precious time. This has always been very much appreciated.

The third person that I want to say thanks to is Doctor Phung Nhu Hai from The Institute of Information Technology (AMST) for his guidance on how to build machine learning models.

Last but not least, I would like to give thanks to my family and close friends, for their support and encouragement throughout my study.

October, 2023

Do Duc Huy

Commitment

Respectfully addressed to:

− Board of Hanoi Medical University

− Board of Institute for Preventive Medicine and Public Health

− Department of Epidemiology

− Board of Dissertation Assessment

My name is Do Duc Huy - Master student in Epidemiology in English of Hanoi Medical University, cohort 2021-2023, hereby declare that: This is a research that I conducted under the scientific guidance of Assoc.Prof Le Minh Giang and Assoc.Prof Nguyen Thi Trang. The data and results presented in the research are completely truthful. In addition, the thesis also uses a number of comments, assessments as well as results from other authors, agencies and organizations, all with source annotations clearly stated in the references. I will take full responsibility if there was any fraud in the contents of my research.

October, 2023

Do Duc Huy

Abbreviation

Abbreviation	Explanation
AI	Artificial intelligence
ANNs	Artificial Neural Networks
AFP	Maternal Serum Alpha-fetoprotein
uE3	Maternal serum Estriol
HCG	Human Chorionic Gonadotropin
PAPP-A	Pregnancy-associated Plasma Protein A
NIPT	Non-invasive Prenatal Screening
KNN	K-nearest neighbor
SVM	Support Vector Machine
RF	Random Forest
MLP	Multilayer Perceptron
XGBoost	Extreme Gradient Boosting

List of Tables

1.1 Diagnosis of abnormalities using AFP, uE3 and HCG 8
1.2 Confusion Matrix 19
2.1 Data collection tools available at each hospital 24
2.2 Extracted information 26
3.1 Characteristics of participants in trimester 1 34
3.2 Characteristics of participants in trimester 2 35
3.3 Model performance in training process on trimester 1 data 42
3.4 Model performance in validating process on trimester 1 data 43
3.5 Model performance in training process on trimester 2 data 44
3.6 Model performance in validating process on trimester 2 data 45
3.7 Models performance in testing process on trimester 1 data 47
3.8 Models performance in testing process on trimester 2 data 49
4.1 Detailed hyper-parameters of each model 57

List of Figures

1.1 Maternal age-related risk for trisomy 21 at 12-week gestation and maternal serum b-hCG levels (left) and PAPP-A (right) 7
1.2 Examples of a supervised learning system 11
1.3 Example of kNN algorithm 12
1.4 Example of Random Forest algorithm 13
1.5 Example of Support Vector Machine algorithm 14
1.6 ROC Curve 21
2.1 Study diagram 26
3.1 Data Overview 31
3.2 Data Overview 33
3.3 Ranking variables importance in trimester 1 ultrasound data 36
3.4 Ranking variables importance in trimester 1 biochemical data 37
3.5 Ranking variables importance in trimester 1 ultrasound and biochemical data 38
3.6 Ranking variables importance in trimester 2 ultrasound data 39
3.7 Ranking variables importance in trimester 2 biochemical data 40
3.8 Ranking variables importance in trimester 2 ultrasound and biochemical data 41
3.9 Comparison of models on trimester 1 testing data through ROC curve . 46
3.10 Comparison of models on trimester 2 testing data through ROC curve . 48
4.1 Demonstration of overfitting 52
4.2 Different type of cups 53

Introduction

Down syndrome (DS) is a congenital defect caused by an extra 21st chromosome[1]. This is the most common chromosomal disorder in the US that appears in 1 in every 700 babies and approximately 250,700 people were living with Down syndrome in the US in 2008[2]. Children with Down syndrome have a higher risk of congenital heart disease[3], deafness[5], ear infections, lung infections[6] and autism[7] leading to high mortality rates and reduced life expectancy. This syndrome not only affects the children themselves, but also causes significant economic and emotional burdens on the family and society. Children with Down syndrome have greater unmet needs than other children with special health care needs[8] and the cost of caring for a child with Down syndrome aged zero to four is four times higher than the costs of caring for a child of the same age without the syndrome[9]. Currently, there is no cure for Down syndrome[1]. Prenatal and early screening is necessary for the early detection of Down Syndrome so that the mother can make appropriate decisions for the baby, including early cochlear implantation, heart surgery, and respiratory support[10]. If a fetus is suspected to have Down syndrome when the mother's age is greater than 35 years[11] or there is thick nuchal translucency detected through ultrasonography[12], screening methods are implemented to calculate the woman's risk of delivering a child with Down Syndrome.

Currently, there are various methods of risk assessment for detecting Down syndrome. Among them, Non-Invasive Prenatal Testing (NIPT) has the highest sensitivity and specificity, with upwards of 99%[13]. However, NIPT is more expensive than other screening methods (ranging from 3 to 6 million Vietnamese Dong) to be adopted as a universal screening program . Alternative methods of detection, including double and triple tests, but with lower sensitivity and specificity, ranging from 50%-60% sensitivity and 85-90% specificity, are being used widely due to their lower costs[14].

In recent years, with the expansion of artificial intelligence (AI) to all areas of

science that use available medical data efficiently to build decision support systems[15], there are also new AI-based methods for the early detection of Down syndrome. Machine learning is the most important part of AI, it gives computer systems the ability to learn automatically. A well-developed machine learning model can achieve sensitivities greater than 95% and higher with more data, higher than current screening methods used for the detection of Down syndrome in Vietnam, which are double and triple tests. For example, Neocleous et al developed a machine learning model with 100% sensitivity for screening Down syndrome based on three different dataset with a total of 122,362 aneuploidies and 967 malformations[16]. Machine learning models can be implemented as a mobile app or a website, so prenatal screening using this method only requires a smart device that can access the software. Unlike NIPT, it is a cheap and easy-to-use method that can be used any where at any time. Therefore, we can apply this screening method in the healthcare system in Vietnam, especially at the commune level, where there are no trained specialists in genetics. This new method will help to increase the rate of pregnant women who can access the screening program and hopefully result in a reduced the frequency of undetected babies with Down syndrome in Vietnam.

There have been a few machine learning models that screen for Down Syndrome in the fetus but they are all built based on non-Vietnamese populations. This study is part of a national study entitled “Research to build an artificial intelligence system to support prenatal screening for some common abnormalities in Vietnam”. It aims to increase the proportion of pregnant women being screened for abnormalities in the community. In the parent study, we developed four models to screen for Down Syndrome in the first and second trimester based on data from pregnant women at Vietnam National Hospital Of Obstetrics and Gynecology.

In the present study, in order to adapt to the real life setting which might not have all the test results required due to limitations at their health care facilities, we used 3 different combinations of variables in each trimester to build prediction mod-

els for Down Syndrome. These included, combinations of ultrasound test results only, a combination of biochemical tests results only, or the combination of ultrasound and biochemical test results. We then compared the result of these models to find the best model with the highest sensitivity and specificity as well as the least amount of information required. The outcome variable was whether the fetus had Down Syndrome as diagnosed by the gold standard test of amniocentesis.

The objectives of this study were:

Objective 1: To describe the process of making machine learning models in prenatal screening for Down Syndrome using data from pregnant women collected from the Vietnam National Hospital Of Obstetrics and Gynecology.

Objective 2: To assess the sensitivity and specificity of machine learning models in prenatal screening for Down Syndrome using data from pregnant women collected from Hanoi Medical University Hospital to find the most appropriate model.

1 Chapter 1: Literature Review

1.1 Down Syndrome

1.1.1 Overview

Down Syndrome, is one of the most common chromosomal conditions. It occurs in about 1 in 800 births worldwide. In the United States, Down Syndrome is found in 500 live births annually and more than 200,000 people are living with the condition. Down Syndrome was first found and described by John Langdon Down, a physician from Cornwall, England. More than 90 years later, the chromosomal cause was delineated and the condition was named Down Syndrome.

The potential for the development and socialization of persons with DS has been increasingly realized, and early support for affected children and their families is widely implemented, although the disparity in access to health care and other supportive resources still exists. There is considerable phenotypic variation among patients, and intellectual disability is most commonly moderate but ranges from mild to severe, whereas social function is often high relative to cognitive impairment. There are also differences in the incidence and presentation of Down Syndrome according to ethnic background and geographic region.

1.1.2 Genetic Features of Down Syndrome

Down syndrome is thought to be caused by an extra chromosome 21 in the genome, also known as trisomy 21 or trisomy 21. Trisomy 21 is the most common inherited chromosomal disorder that occurs when a child is born with an extra copy of chromosome 21 during pregnancy. In all cases of normal reproduction, both parents pass on genes to their offspring, which are carried in chromosomes. As the baby's cells grow, each cell must receive 23 pairs of chromosomes from the mother and 23 from the father, for a total of 46 chromosomes. In children with Down syndrome, one of

the chromosomes does not separate properly. Babies end up with three copies, or one extra copy of chromosome 21, instead of two. Thus, the fetus has 47 chromosomes due to an extra chromosome number 21. It is this "excess" chromosome that disrupts normal physical and intellectual development, causing physical and mental disorders.

1.1.3 Prenatal screening methods

1.1.3.1 Ultrasound Ultrasound is a non-invasive procedure that does not harm both the mother and the fetus, which allows clinicians to gather some information about the pregnancy that cannot be provided by any examination such as gestational age, number of fetuses, fetal development, mother-to-child metabolism quality (based on Doppler) and fetal morphology. Although there have been many technical improvements, ultrasound is still not the perfect method, it can only detect some fetal malformations when the fetus is in a favorable position with the right amount of amniotic fluid. Unclear morphological abnormalities are also difficult to detect on ultrasound. In Down Syndrome, ultrasound can only detect indirect images such as nuchal translucency.

Ultrasound for the measurement of nuchal translucency is usually done at 11-13 weeks of pregnancy which will give the most accurate results. The majority of cases with nuchal translucency < 3 mm were classified as low-risk (less likely to develop chromosomal abnormalities). In the case when nuchal translucency is ranged from 3.5 to 4.4 mm, there is a chromosomal abnormality rate of 21.1% and in the case which it is >= 6.5 mm, the risk of chromosomal abnormality can be increased up to 64.5%. In cases where nuchal translucency is > 3 mm, the pregnant woman will be ordered to perform an additional triple test at 16-18 weeks.

1.1.3.2 Double test, Triple test The first Down syndrome screening method was introduced in the 1970s based on maternal age. women over 40 years old will be given an amniocentesis test to determine the risk of fetuses with chromosomal ab-

normalities. Later, when amniocentesis becomes safer than before with the guidance of ultrasound, the cost is also reduced, amniocentesis is widely indicated in high-risk pregnant women, ie older than or equal to 35 years old.

- **Maternal serum alpha-fetoprotein (AFP)**

A developing fetus has 2 main types of blood protein, Albumin and alpha fetoprotein (AFP) while an adult has only albumin, so an AFP test in the maternal serum is used to indirectly determine the amount of AFP in the fetal blood.

Normally only a small amount of AFP in the amniotic fluid can cross the placenta to enter the mother's bloodstream. However, when there is a neural tube abnormality, because part of the embryonic neural tube is not closed, AFP will escape into the amniotic fluid. Neural tube abnormalities include anencephaly (due to the neural tube that does not close the head) and spina bifida (due to the inability of the tail of the neural tube). In the US, the rate of these diseases is 1-2/1000 births. Likewise, in gastroschisis or omphalocele, AFP from the fetus enters mother's bloodstream in a larger amount than usual.

AFP tends to be lower than normal in fetuses with Down syndrome or some chromosomal abnormalities, so AFP is useful in screening for Down syndrome and several other infections. A combination of AFP screening and ultrasound can detect almost all anencephaly and most cases of spina bifida.

- **Maternal serum free Beta-HCG**

This is the most commonly used test during pregnancy. About 1 week after the embryo implants in the uterus, the amount of beta HCG secreted by the culturing cells is sufficient to diagnose pregnancy. In the early stages of pregnancy, beta HCG helps in early diagnosis and prognosis of miscarriage, ectopic pregnancy because in these cases, beta HCG is lower than normal.

Later in pregnancy, at the end of the second trimester, HCG may be used in combination with AFP to screen for specific chromosomal abnormalities in Down syndrome. Increased HCG in association with decreased AFP is an implication of Down syndrome. Meanwhile, abnormally high hCG suggests pseudocyesis.

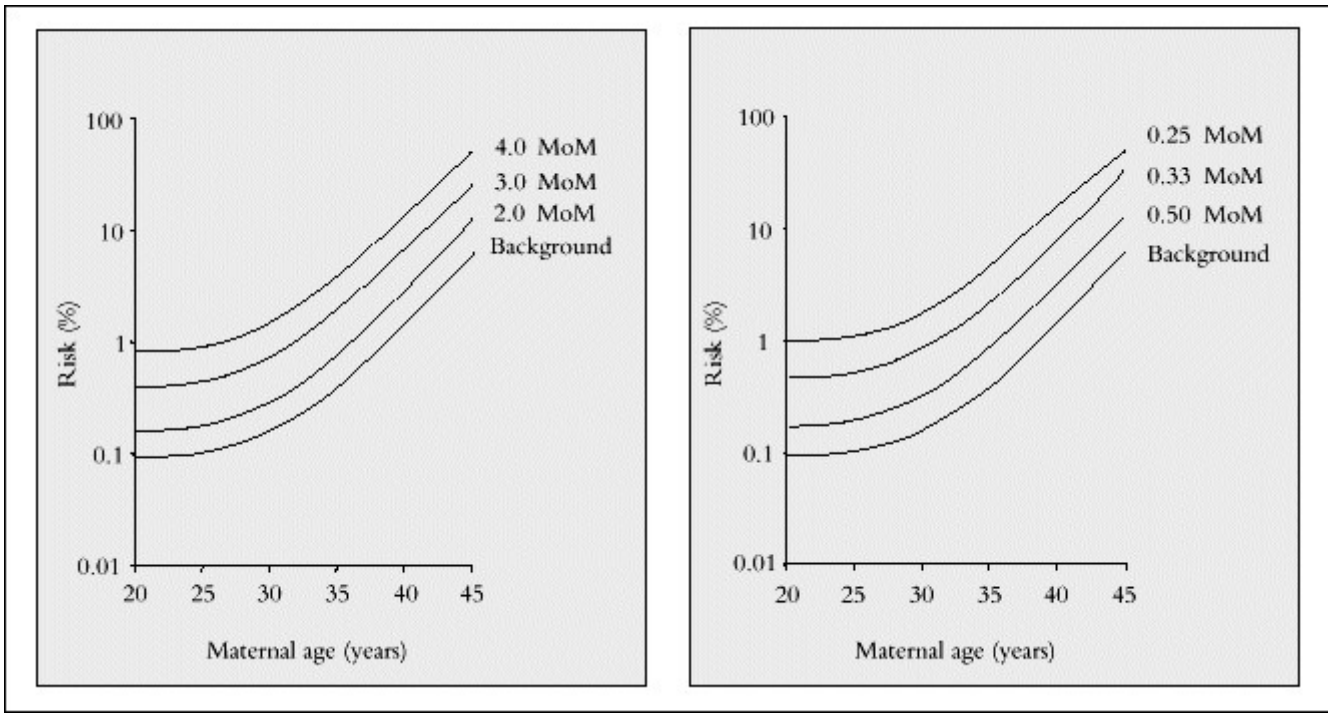

Figure 1.1: Maternal age-related risk for trisomy 21 at 12-week gestation and maternal serum b-hCG levels (left) and PAPP-A (right)

- **Maternal serum Estriol (uE3)**

Estriol is derived from dehydroepiandrosterone (DHEA) which is produced from the adrenal glands and then converted to estriol by the placenta. Estriol enters the mother's bloodstream and is excreted in the urinary tract or excreted by the liver into the bile. Continuous testing of estriol in the third trimester is performed to monitor fetal health status. If the concentration of estriol is reduced, the fetus is at risk and may indicate an end to pregnancy. Estriol is also reduced in fetuses with Down syndrome or adrenal insufficiency or anencephaly.

- **Pregnancy-associated plasma protein A (PAPP-A)**

In the first trimester, low serum PAPP-A is an indication of trisomies 13, 18 and 21. Furthermore, low PAPP-A levels in the first trimester predict a low birth weight pregnancy or stillbirth. A higher than normal PAPP-A suggests a larger than normal fetus. A combination of serological tests can potentially increase the sensitivity and specificity of detecting fetal abnormalities. The classic 3 screening tests includes alpha-fetoprotein (MSAFP), beta-HCG, and estriol (uE3). Some facilities use fourth tests, which is inhibin-A.

Table 1.1: Diagnosis of abnormalities using AFP, uE3 and HCG

Syndrome	AFP	uE3	HCG
Neural tube defects	High	Normal	Normal
Trisomy 21	Low	Low	High
Trisomy 18	Low	Low	Low

1.1.3.3 Non-invasive prenatal screening (NIPT) This is considered to be the most effective and safest testing method available today. The method is performed early from the 10th week of pregnancy through the mother's blood sample (only 7-10 ml). Chromosome abnormalities can be screened including chromosome 6, 9, 13 (Patau's syndrome), chromosome 18 (Edwards), chromosome 21 (Down), chromosome X, Y, and segmental mutations, etc. In addition, this method is also applicable for single pregnancy, twins, surrogacy with high accuracy, up to 99.98%.

1.1.4 Diagnosing Down Syndrome

1.1.4.1 Amniocentesis (golden standard of this study) Amniocentesis is the most widely used method today because of its technical simplicity as well as a low rate of complications. It is considered the main method of obtaining fetal specimens.

Amniocentesis is done in 3 periods: Early amniocentesis (13 to 16 weeks gestation), classic amniocentesis (from 17 to 20 weeks gestation), late amniocentesis (after 20 weeks).

The best gestational age for this procedure is 17 to 18 weeks because at this time the chance to successfully draw out amniotic fluid is highest while the rate of complications for both mother and fetus is lowest. The procedure is performed under ultrasound guidance.

1.1.4.2 Chorionic Villus Sampling Chorionic villus sampling (CVS), or chorionic villus biopsy, is a prenatal test that involves taking a sample of tissue from the placenta to test for chromosomal abnormalities and certain other genetic problems. This method causes a high rate of miscarriage (about 9%), so it is only used mainly in cases of fetuses with severe abnormalities detected in the first trimester. This method is performed under ultrasound guidance. Results will be available after 5 to 7 days.

1.2 Artificial Intelligence (AI)

1.2.1 Definition

Artificial Intelligence, also known as AI, is the intelligence expressed by machines, different from the natural intelligence expressed by humans or animals, which are related to consciousness and emotions. In other words, artificial intelligence is a branch of computer science whose purpose is to give the software the ability to analyze information, then make decisions based on results.

1.2.2 Machine learning

Machine learning is the most important part of AI, it gives computer systems the ability to learn automatically. Machine learning can be understood as the process by which a system “learns” itself from past experiences and converts those “lessons” into its knowledge, instead of using knowledge obtained by humans.

Currently, two definitions of machine learning method have been proposed. The first definition by Arthur Samuel describes this approach as “the field of study

that gives computers the ability to learn without being explicitly programmed." This is the old and unofficial definition. Instead, Tom Mitchell proposed a more modern definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P improves with E". For example, in computer-based chess, where E is the experience from playing previous games, T is the task of playing chess, and P is the probability that the software will win the next time it is played.

Any machine learning method can be categorized into one of two types: supervised learning or unsupervised learning.

1.2.2.1 Supervised learning As children, we learned to classify new things under the guidance of adults. They point at a furry four leg creature that bark and tell us that it is a dog. Through many such instructions, we get to know how to identify a dog among other animals. This is the core concept of supervised machine learning. In this method, the software is given a dataset and already knows what the correct output should look like, having the idea that there is a relationship between the input and the output. The task of a supervised machine learning mechanism is to try to find the relationship between the input data and the desired output, and then use that relationship to predict the output for the new input.

The advantage of the supervised machine learning method is that it can find out the correlation between input and output data that is close to reality and has good coverage of different cases. However, the disadvantage of machine learning methods is that large input data is required, and the data must be pre-labeled, which can be expensive in terms of time and money. In addition, in order to have good coverage of the different cases, the input data must be diverse and need to be updated continuously, because although it is called a cat, the cats in different places are different in shape, size, color, and sometimes we have to ask ourselves if it's a cat, which is the same in this machine learning approach.

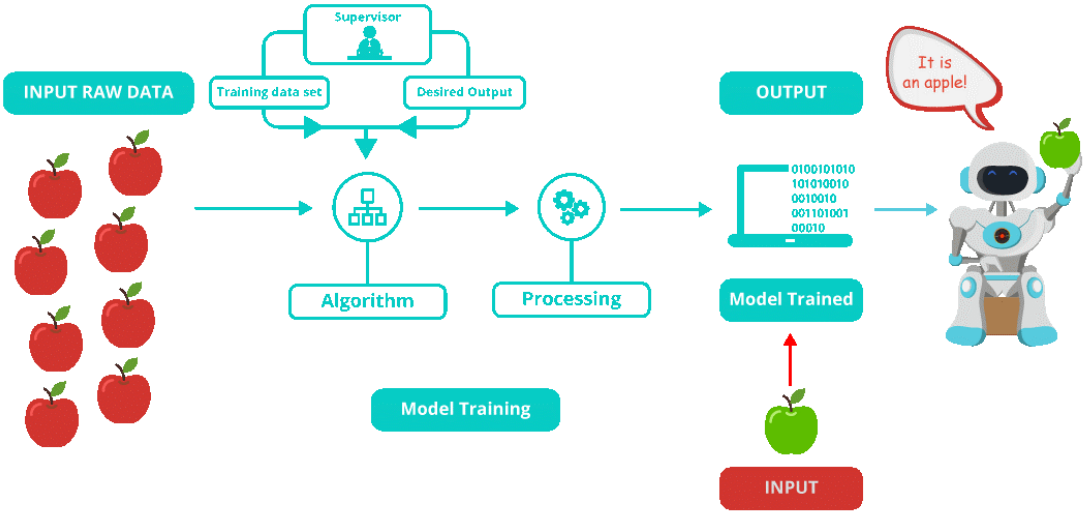

Figure 1.2: Examples of a supervised learning system

1.2.2.2 Unsupervised learning Unsupervised machine learning is used when we know little or not what the outcome will be. While in supervised machine learning, we try to build a predictive model based on labeled input data, in the unsupervised learning method, instead of available labeled data, the model collects data from the environment and labels them itself, just like children who can imitate the actions of adults, classifying other animals by themselves, or come up with rules of a game based on observations.

The advantage of the unsupervised machine learning approach is that it does not need labeled data, and it can provide unknown information from the input data, as well as automatically classify the data by finding different characteristics from the data itself. However, this method has disadvantages such as it takes many steps to build, and it is difficult to understand what is going on inside the software or what method it is using to learn.

1.2.3 Machine learning algorithms

1.2.3.1 K nearest neighbor (kNN) K nearest neighbor (or kNN) is one of the simplest supervised machine learning algorithms mostly used for classification. It classifies an observation based on how its neighbors are classified.

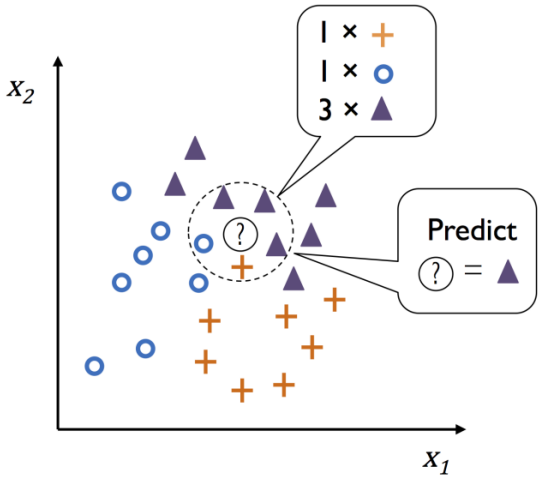

Figure 1.3: Example of kNN algorithm

Figure 1.3 presents an example of kNN algorithm with 3 different categories (or "class" in machine learning language). In this example, there are the most observations in the 3rd group in the area we call "neighbor" (k in kNN is the number of observations in this area so the more k the larger the area) so we can predict that the unknown observation belongs to the 3rd group. Thus, kNN algorithm is easy to use and it adapts easily to new data, but its accuracy falls when we need a large number of variables to classify.

1.2.3.2 Random forest (RF) Random forest (RF) is a method that operates by constructing multiple Decision Trees during training phase. The Decision of the majority of the trees is chosen by the random forest as the final decision.

Figure 1.4 shows an example of RF algorithm in classifying if this fruit is cherry or orange. There are multiple decision trees, each with its own parameters and conditions. In the end, the majority of decision trees classify that this fruit is orange

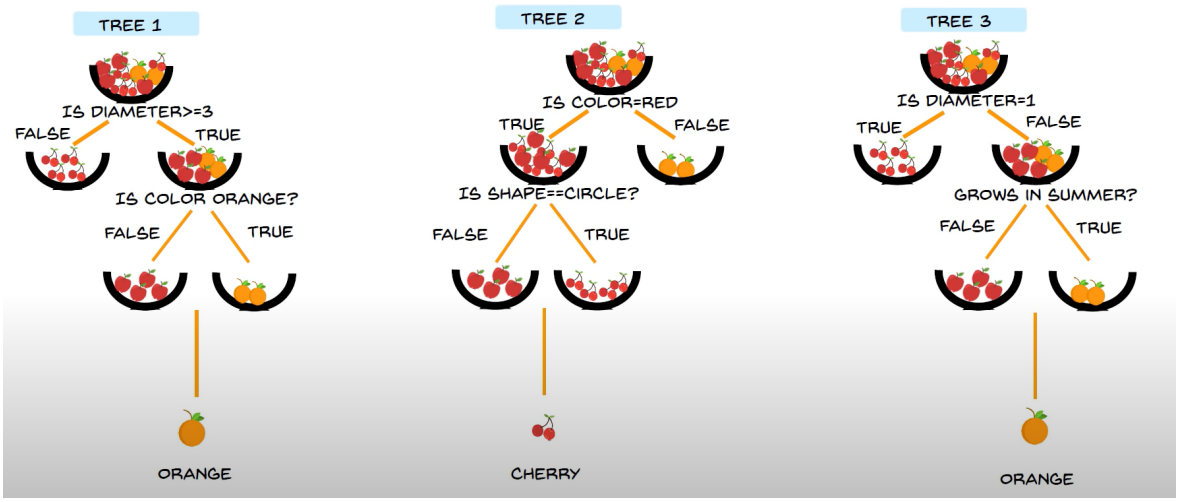

Figure 1.4: Example of Random Forest algorithm

so the final conclusion of this algorithm is orange. Because of many decision trees with different parameters and conditions, RF can maintain accuracy when there is missing data but it is also more complex and requires more computing power than other machine learning algorithms.

1.2.3.3 Support vector machine (SVM) Support Vector Machine (SVM) algorithm construct a hyperplane (red line in figure 5) where the distance between two groups of data points is at its maximum. This hyperplane is known as the decision boundary, separating the groups of data points (e.g., oranges vs. apples) on either side of the plane.

1.2.3.4 Extreme Gradient Boosting (XGBoost) Extreme Gradient Boosting algorithm is an implementation of gradient boosted decision trees. Both XGBoost and RF combine the outputs from individual trees but these two algorithm differ in the way individual trees are built and in the way the results are combined. In RF, we build independent decision trees and combine their result in parallel while XGBoost combines the result sequentially so that each new tree corrects the error of the previous tree. The first step is to fit a single decision tree then evaluate how well this tree

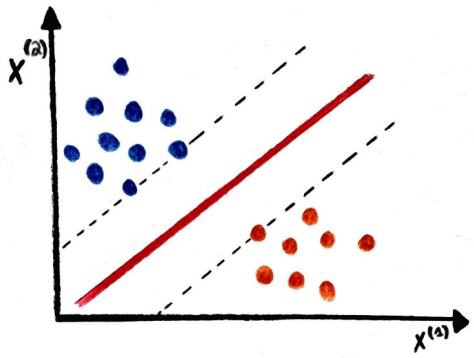

Figure 1.5: Example of Support Vector Machine algorithm

does in classifying the data. Then we add the second tree that its performance better than the first tree alone. Then we keep adding trees until a point that when we add new trees, the performance of the whole system stops increasing.

1.2.4 AI application in health care

In the clinical setting, AI is often used to build clinical decision support system. Clinical decision support systems have the responsibility of assisting physicians or healthcare professionals in making clinical decisions. These systems can improve the quality of healthcare by providing references to doctors based on data from the past.

Clinical decision support systems have been developed since the 1970s, such as the MYCIN, INTERNIST-1 and CASNET. Decision support systems have been developed and applied in many fields of healthcare such as diagnosing diseases, developing treatment regimens, making drugs, monitoring and taking care of patients, … These systems have been applied in many facilities around the world. AI algorithms in healthcare have been developed by many companies, including large companies such as IBM, Microsoft, Google, Intel, Facebook and even startups. In recent years, the development and application of clinical decision support systems have been increasingly strengthened. There are many reasons for this development, that is the

development of hardware leading to increased computing power, shortening the time to collect and process data; the volume of medical-related data collected from medical and personal devices increasing; development of genetic databases; electronic medical record management systems are becoming more and more popular; development of highly accurate advanced AI techniques (such as deep learning methods) in the fields of computer vision and natural language processing enhances the accuracy of AI system.

1.2.5 Studies applying AI in prenatal screening for Down syndrome

The Fetal Medicine Foundation's algorithm is being used widely in prenatal screening for Down Syndrome. Many studies were conducted to measure the efficiency of this software. A study conducted by Kevin Spencer in a population of 30 cases of Down syndrome and 11758 unaffected pregnancies concluded that risks produced by the Fetal Medicine Foundation agree very closely with Down syndrome prevalence with the correlation coefficient being 0.9995[17]. Another study by Stephen P.O. et al on 8 cases of Down Syndrome in 1000 fetuses yielded 75% sensitivity[18].

Not only test AI models that already built, many studies use novel methods to build their own models. Falin He et al built an AI software on 58923 negative Down syndrome cases and 49 Down syndrome cases and test this software on a dataset of 27,143 negative Down syndrome cases and 27 Down syndrome cases from other hospital, which achieved 85.2% sensitivity and 95% specificity[19].

Neocleous used artificial neural networks to assess the risk of multiple aneuploidies, Down syndrome and other aneuploidy mutations[16,20]. The authors introduced a non-invasive diagnostic procedure for fetal malformations in early pregnancy, by proposing a method using an artificial neural network trained with data from single pregnancies in first-trimester screening. Three different dataset with a total of 122,362 euploids and 967 malformations were used. Data for each case contained markers collected from the mother and fetus. The proposed artificial neural network models have

been optimized in the sense of achieving a minimum false positive rate and at the same time guaranteeing a 100% sensitivity for Down syndrome. These systems also accurately identify other malformations (Edwards, Patau, Turner and Triploid syndromes) achieving sensitivity greater than 80%. The results of the study demonstrate that artificial neural network systems can support effective, non-invasive early screening for fetal malformations with better outcomes than other currently available methods.

Some other studies applied techniques such as Fuzzy Cognitive Maps, logistic regression, a two-stage approach to diagnose the risk of Down syndrome and neural tube defects[21–23].

Catic et al used two neural network architectures to classify five prenatal syndromes (Turner, Klinefelter, Patau, Edwards, and Down) based on maternal serum screening test results. ultrasound and patient demographics[24]. The purpose of this work is to test the effectiveness of different neural network architectures for this task. This study demonstrated that relatively simple neural network architectures, such as feedforward, can have high classification accuracy. Because of the non-linear input-output relationship, the classification accuracy can be better achieved with a recursive neural network architecture, such as the Elman neural network. The feedforward neural network with 15 neurons in the hidden layer achieved a classification sensitivity of 92.00%. The classification sensitivity of the Elman neural network is 99.00%. The average accuracy of the feedforward neural network is 89.6% and the response is 98.8%. Koivu et al studied, tested, and evaluated machine learning algorithms to improve the performance of Down syndrome screening during the first trimester of pregnancy. Machine learning algorithms pose an adaptive alternative to developing better risk assessment models using existing clinical variables[25]. The best performing deep neural network model gave an area under the curve of 0.96 and detection rate of 78% with 1% false positive rate with the test data. Support vector machine model gave area under the curve of 0.95 and detection rate of 61% with 1% false positive rate with the same test data.

Li et al. propose a cascading machine learning framework designed to predict Down syndrome based on three additional stages: 1) pre-judgment with isolation forest technique, 2) model ensemble by voting strategy, and 3) final judgment using logistic regression approach. The test results show that the performance of this framework on the maternal serum screening dataset, when evaluated with different evaluation parameters, outperforms some machine learning methods[26]. The best suggested combination of input features for Down screening are alpha-fetoprotein (AFP) group, human chorionic gonadotropin (hCG), unconjugated estriol (uE3), and maternal age. In addition, the method proposed by the authors is capable of producing even more accurate predictions for the data.

1.3 Evaluating the effectiveness of artificial intelligence software

1.3.1 Dataset and model evaluation

Model evaluation is usually performed on data that the model has never been learned from – on validation sets and test sets. Different problems will have different evaluation criteria. A dataset has 3 functions: model training, model fine-tuning and model evaluation. The dataset also has 3 parts: training set, validation set and test set.

Machine Learning models are trained based on data. The more “good” data is fed into the model, the more accurate the model’s predictions. Therefore, most of the data in the dataset are used to train the model, and this piece of data is called the training set. On the other hand, there is a need for a piece of data to evaluate the “learning” of the model, then we need a test set. The data in the test set should be new, never been “learned” or “seen” by the model, and close to the actual data. The test set acts as a sample with population being the actual data of the problem being solved, or the data in the test set having the same distribution as the actual data. Therefore, the model results shown on the test set are a reliable measure of the actual performance of the model. Assuming you want to recognize license plates in night shots, the test set

should now be pictures were taken at night (same distribution with actual data) instead of those taken in good lighting conditions. At this time, the results of the model on the test set will be more reliable when the test set contains an image that is not close to the actual problem to be solved.

During model training, there will be cases where the model is overfitting with the training data. Simply put, overfitting is a phenomenon where the model performs well when evaluated on the training set but does not perform well on the test set. Therefore, we need to calibrate the model (fine tune the hyper parameters) to improve the evaluation results on the test set. At this point, we need a dataset that the model has never been "learned" or "seen" before, for "early warning" about problems that the model may encounter when exposed to the data. reality (like underfitting or overfitting). So this dataset must be distributed with test set or actual data, and it is called validation set or development set. A question arises: Is it possible to use the test set as a validation set? The use of test sets to calibrate the model is not recommended in practice in order to keep the "unseen" nature of the test set. If the model is improved from the test set in order to "fit" better with the test set itself, the model tends to overfit with the test set, reducing the reliability of the evaluation results on the test set.

1.3.2 Evaluation parameters

1.3.2.1 Accuracy Accuracy (accuracy) simply measures how often the model correctly predicts. Accuracy is the ratio between the number of correctly predicted data points and the total number of data points:

$$accuracy = \frac{number\ of\ corrected\ predictions}{number\ of\ total\ data\ points}$$

However, a model with high accuracy is not necessarily good. Accuracy exposes its limitations when used on an unbalanced dataset. We have the following example:

We want to screen for Down syndrome in fetuses of pregnant women. We can collect data from 10,000 pregnant women but only 10 cases with Down syndrome fetuses. It is easy to see that, as long as the model always predicts that all pregnant women are normal, the model has an accuracy of 99.9%. However, in practice your model cannot detect pregnant women with Down. Therefore, our dataset is imbalanced, so relying on accuracy to evaluate the model does not bring many positive results.

1.3.2.2 Confusion Matrix (2x2 table) The disadvantage of Accuracy is that it only tells us the accuracy of the model's prediction, but it does not show how wrong the model is predicting, so we need another evaluation method - Confusion Matrix. Confusion matrix is a technique for evaluating model performance for classification problems. Confusion matrix is a matrix that represents the number of data points that belong to a class and are predicted to belong to the class.

In order to classify whether a pregnant woman has a fetus with birth defects, we have to answer a yes or no question, or in other words, positive or negative. There are 4 possibilities when comparing the software's prediction with fetus's true condition. If the prediction says that this case is positive and in fact this person is positive, this is called a true positive, but if in fact this person is negative, it is called a false positive. Conversely, true negative occurs when both the prediction and fetus's true condition are negative, and false negative occurs when the prediction is negative when in fact it is positive. We can draw this table from the explanation above:

Table 1.2: Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

The sensitivity (sometimes also named the detection rate in a clinical setting) of the software is the proportion of fetuses which test positive for birth defects among

those which truly have the condition. Mathematically, this can be expressed as:

$$\text{Sensitivity} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Specificity of the software is the proportion of fetuses which test negative for birth defects among those which truly do not have the condition. Mathematically, this can also be written as:

$$\text{Specificity} = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Finally, accuracy is the combination of true positive and true negative cases among the total population. Mathematically, this can also be written as:

$$\text{Accuracy} = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

1.3.2.3 ROC curve In classification problems, classification algorithms often predict the score or probability of belonging to a class of input data. This helps us to know the certainty of the model when classifying. After predicting probabilities or scores, it is necessary to convert those values to the labels of the classes. The transition from probabilities, scores to labels is determined by a threshold. The output of Logistic Regression is the value of the sigmoid function:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

The value of $\sigma(x)$ is in [0,1], representing the probability that the input data point belongs to the positive class. To convert the probability to the class label, we need to determine the threshold value. The default threshold value is 0.5, which means:

If S(x) > threshold (0.5), the output of the model is 1

If S(x) < threshold (0.5), the output of the model is 0

The problem is that sometimes default threshold = 0.5 is not the best classifier "threshold", this happens when the classes of the problem are not balanced (predicting a rare disease with extremely low probability), or the priority of one type of error is higher than the other, and so on. Therefore, sometimes we need to change the threshold for the model to achieve the desired results. ROC curve is a tool to choose the appropriate threshold for the model. For each threshold value, we obtain two values represented on the ROC curve:

- True Positive Rate (TPR or Sensitivity - Recall): is the sensitivity of the model, indicating how accurate the prediction is in the positive class. TPR is the quotient of the number of correctly predicted data points in the positive class with the number of data points in the positive class.
- False Positive Rate (FPR): is the probability of getting Type II Error Where, Specificity indicates the accuracy of the prediction in the negative class.

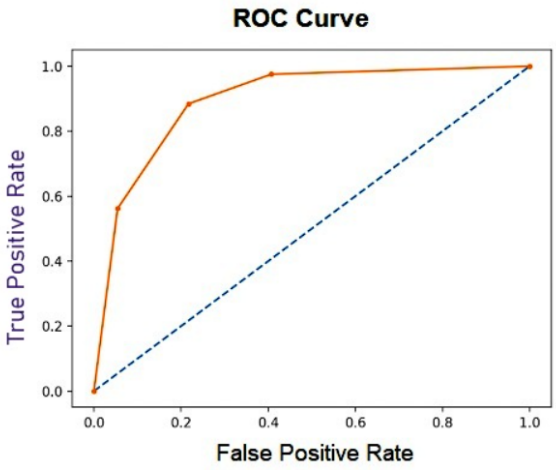

Figure 1.6: ROC Curve

The orange points represent each threshold, where the vertical axis is the TPR value and the horizontal axis the FPR value. Connect the orange points to get the ROC

curve. The blue dashed line represents the results of the “no skill model” – the model predicts by randomizing the results. It should be noted that the lower the FPR value, the lower the probability of having a Type II Error, so the points on the left should be considered more if we need to minimize False Negative (Type II Error). On the other hand, the higher the points lie, the larger the TPR. Depending on the problem to choose a point – corresponding to a suitable threshold.

2 Chapter 2: Study Subjects And Research Methodology

2.1 Description of the parent study

The parent study, entitled “Research to build an artificial intelligence system to support prenatal screening for some common abnormalities in Vietnam”, was a cross-sectional study conducted to build expert systems and machine learning models to screen for four abnormalities in Vietnam: thalassemia and three trisomies which were Down, Edward, and the Patau syndrome.

2.1.1 Study timeline

This study was conducted at the Vietnam National Hospital Of Obstetrics and Gynecology and Hanoi Medical University Hospital from September 2020 to March 2023.

2.1.2 Study participants

Data were collected from the medical records of pregnant women who visited the National Hospital of Obstetrics and Gynecology from January 2012 to December 2022 were included. In order to test built models using external data, the medical records of pregnant women examined at Hanoi Medical University Hospital, Hanoi Obstetrics and Gynecology Hospital and the National Hospital of Obstetrics and Gynecology were collected.

Eligible participants were pregnant women who had either ultrasound test results, total peripheral blood cell analysis, prenatal screening test results (double or triple test), serum iron and ferritin test results, hemoglobin electrophoresis test results and chromosome and/or thalassemia gene test results (in high-risk cases). We excluded those with multiple pregnancies or IVF pregnancies due to their differences in

ultrasound and biochemical test results compare to those without these conditions.

2.1.3 Sample size and sampling methods

We used convenience sampling methods to collect data. We collected data from 9845 records of pregnant women from Vietnam National Hospital Of Obstetrics and Gynecology and used this dataset to build machine learning models.

2.1.4 Variables used to build machine learning models

A total of 16 variables were used to build machine learning models. These included 2 maternal characteristics: mother's age, history of having children with Down Syndrome, 2 double test indices: MoM-hcgb and MoM-papp-a, 3 triple test indices: MoM-ue3, MoM-afp and MoM-hcg, and 9 ultrasound test indices: gestational age, nuchal translucency, fetal crown-rump length, biparietal diameter, fetal heart rate, head circumference, abnormal nose (yes; no), abnormal fetal heart (yes; no), femur length. The outcome variable is whether the fetus was at risk for having Down Syndrome as assessed by clinical experts.

2.1.5 Data collection tools used

Table 2.1: Data collection tools available at each hospital

	National Hospital Of Obstetric and Gynecology	Hanoi Medical University Hospital	Hanoi Obstetrics and Gynecology Hospital
Ultrasound	- Voluson E6 (GE) - Samsung HS60 - Samsung A80 (Samsung)	- Voluson S10 (GE)	- Voluson E6 (GE) - Samsung HS60 (Samsung)
Double/Triple test	Autodelfia (PerkinElmer)	Immulite 2000 (Seimens)	

Table 2.1 shows the testers used for each type of tests in each hospitals.

2.1.6 Potential source of biases and ways to prevent their occurrence

2.1.6.1 Source of biases There were two types of bias that could occur in the present study. Firstly, there was selection bias while selecting which medical record to be included in the study. Second, information bias appeared in 3 stages: Doctors collected false information from pregnant women and wrote them in the medical record and poor quality machines gave wrong measurement and screening results, or different machines from different facilities from different producers with different quality and scales of measurement. During data extraction from medical records, researchers did not fully understand all the information. And finally the third stage, during data input to the database, researchers could incorrectly enter data into our database.

2.1.6.2 Ways to prevent biases Based on these potential biases, we included all medical records that satisfied our study inclusion and exclusion criteria, we used standardized machines for measurement in both hospitals, standardized data input forms, and double checking during data extraction and data input.

2.1.7 Ethical issues

This research was approved by the Institutional Review Board of the Vietnam National Hospital Of Obstetrics and Gynecology, decision number 1776/QĐ-PSTW 29th December 2020. All data collected were inputted into our web-based tool and stored there. Accounts to access the tool were provided to researchers who inputted the data.

2.2 Description of the present study

The present study used data collected from the parent study to build and test four AI models in prenatal screening for Down Syndrome with the outcome variable being whether the fetus has Down Syndrome by the results of the Amniocentesis test.

2.2.1 Study diagram

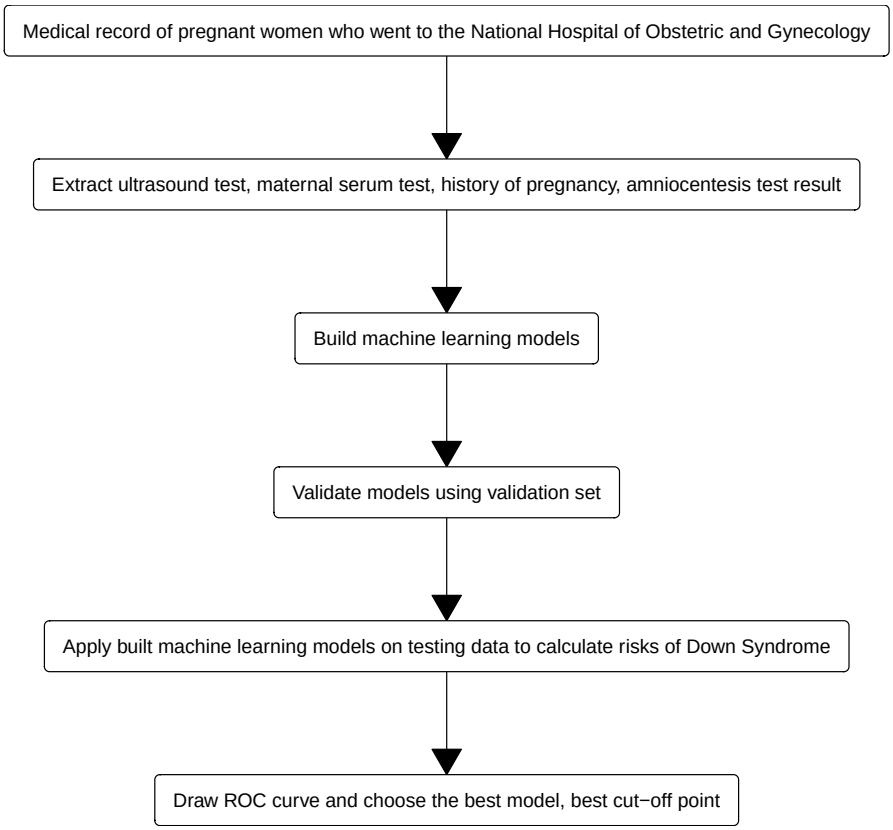

Figure 2.1: Study diagram

All the collected and cleaned study related data were used to build machine learning models. We subsequently, validated these models using a validation dataset, then all built models were tested using a test set, a dataset that were not part of the training data.

2.2.2 Extracted information

Table 2.2: Extracted information

Variable	Data type	Value
Socio-economic status		

Continued on next page

Table 2.2: Extracted information (Continued)

Date of birth		Date
Pregnancy history		
History of having fetus with Down syndrome	Binary	1. Yes; 0. No
Maternal serum		
Free Beta-hCG	Continuous	Unit: MoM
PAPP-A	Continuous	Unit: MoM
AFP	Continuous	Unit: MoM
hCG	Continuous	Unit: MoM
uE3	Continuous	Unit: MoM
Ultrasound test		
Gestational age	Discrete	Unit: Week
Fetal crown-rump length	Continuous	Unit: mm
Nuchal translucency	Continuous	Unit: mm
Biparietal diameter	Continuous	Unit: mm
Fetal heart rate	Continuous	Unit: Times per min
Head circumference	Continuous	Unit: mm
Abnormal fetal heart	Binary	1. Yes; 0. No
Abnormal nose	Binary	1. Yes; 0. No
Femur length	Continuous	Unit: mm
Down Syndrome confirmation		

Continued on next page

Table 2.2: Extracted information (Continued)

Having Down Syndrome by Amniocentesis test	Binary	1. Yes; 2. No

2.2.3 Statistical analysis

Data was analyzed using Rstudio version 4.2.2. All 4 machine learning models were built using caret package version 6.0-94.

Quantitative data was presented as numbers and percentages. Continuous variables were summarized using means, standard deviations, medians, and inter-quantile ranges. The detailed process of building machine learning models can be found in the results section.

2.2.3.1 Assessing sensitivity, specificity of AI models The amniocentesis test is a diagnostic test used to confirm if a fetus has Down Syndrome and this test result is considered to be the "gold standard" of our study. All AI models' results will be compared with the findings from this test result. Four AI models were developed using four machine learning algorithms. They are the K-nearest neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF) and Extreme Gradient Boosting (XGBoost). Each of these algorithms were well-known for their ability in classification problem. The sensitivity (detection rate) and specificity (1 – false positive rate) of each cut-off point in our four AI models were assessed and visualized using Receiver Operating Characteristic (ROC) curves. Sensitivity and specificity were calculated as below:

$$\text{Sensitivity} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{Specificity} = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

For this study, true positives (TP) and true negatives (TN) were the correct

predictions for patients' Down syndrome status, while false positives (FP) and false negatives (FN) were erroneous Down Syndrome predictions. A "false positive" was characterized as the prediction of a pregnant woman carrying a Down Syndrome fetus when she is not, whereas a "false negative" was characterized as the prediction of a pregnant woman not carrying a Down Syndrome fetus when she actually is. The cut-off point for each test result used to classify whether a case had high or low risk of having Down Syndrome will be chosen based on sensitivity and specificity at that point. An optimal cut-off is the point that has the highest sensitivity and highest specificity.

3 Chapter 3: Results

3.1 The process of making machine learning models

3.1.1 Data overview

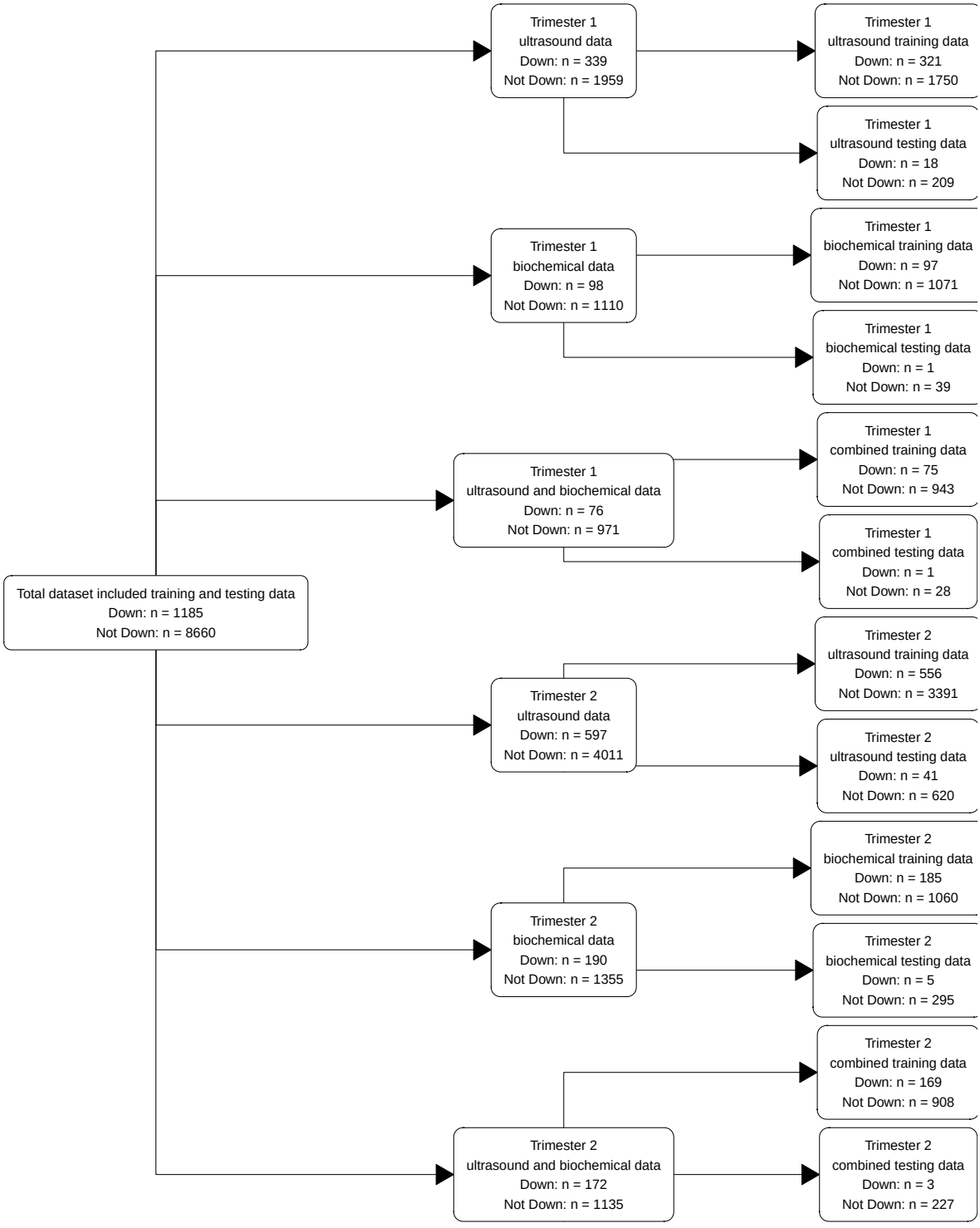

Figure 3.1: Data Overview

Figure 3.1 shows an overview of the dataset and how the initial dataset was broken down into sub-datasets that are used to build different Down-screening modules. The original dataset included 1185 cases of Down Syndrome and 8660 cases without Down Syndrome. This dataset was then split into 6 different datasets for 6 different modules including 2 trimesters (first and second) and 3 modules for each trimester which included data from ultrasound results only, biochemical test results only, and the combination of both ultrasound and biochemical test results. Selected cases in each dataset should have all the corresponding variables which means one case in the trimester 1 biochemical dataset must not have missing data in any of the variables of nuchal translucency, β-hCG or PAPP-A. In order to build machine learning models and test them, each of these 6 datasets were divided into training sets for model training, a validation set for model validation and a testing set for model testing. In this figure, the "training data" included the training and validation set, 80% of each "training data" was the training set and 20% belonged to each test set. The ultrasound dataset were larger than the biochemical dataset from both trimesters. There was more data from trimester 2 than trimester 1.

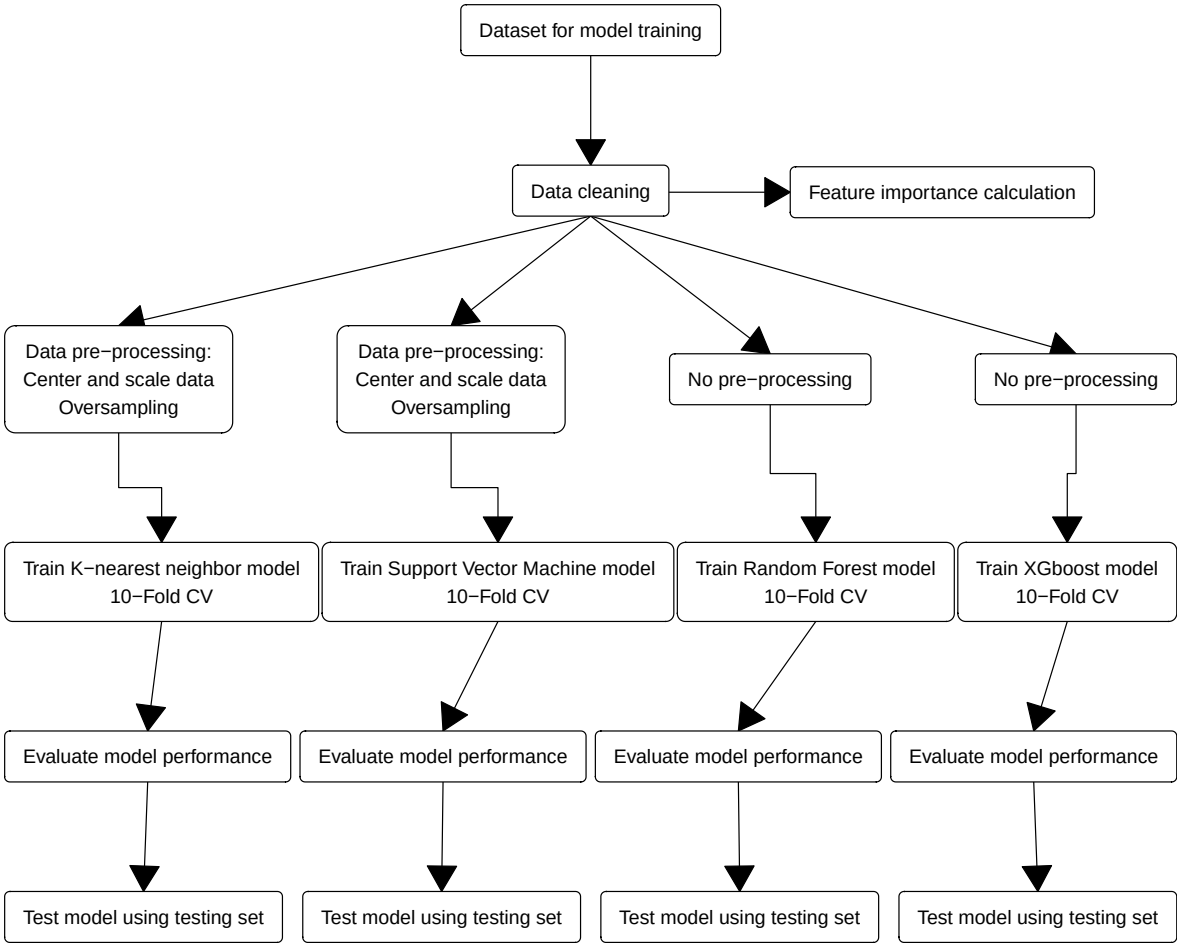

Figure 3.2: Data Overview

All 6 datasets were cleaned and then used to calculate feature importance in order to find which variable was the most important in predicting the occurrence of Down Syndrome. After that, data were centered, scaled and an oversampling method was used before training kNN and SVM models. All RF and XGBoost models did not require pre-processing. Each model was trained using 10-fold cross-validation (90% of data used for training and 10% were used for validating). Finally, after evaluating performance, each of the 24 models were tested using the test set.

3.1.2 Characteristics of participants in each dataset

Table 3.1: Characteristics of participants in trimester 1

	Training		Testing		Overall	
	Not Down	Down	Not Down	Down	Not Down	Down
	(N=2058)	(N=409)	(N=244)	(N=28)	(N=2302)	(N=437)
Mother's age (years)						
Mean (SD)	30 (± 5.7)	32 (± 6.4)	30 (± 5.6)	31 (± 6.1)	30 (± 5.7)	32 (± 6.4)
Mother's age						
≥ 35	510 (25.1%)	158 (38.6%)	51 (21.0%)	8 (28.6%)	561 (24.6%)	166 (38.0%)
< 35	1523 (74.9%)	251 (61.4%)	192 (79.0%)	20 (71.4%)	1715 (75.4%)	271 (62.0%)
Fetus's age (weeks)						
Mean (SD)	89 (± 4.4)	90 (± 4.4)	89 (± 4.4)	89 (± 3.6)	89 (± 4.4)	90 (± 4.4)
History of having children with Down syndrome						
No	2040 (99.1%)	405 (99.0%)	242 (100%)	28 (100%)	2282 (99.2%)	433 (99.1%)
Yes	18 (0.9%)	4 (1.0%)	0 (0%)	0 (0%)	18 (0.8%)	4 (0.9%)
Fetal crown-rump length (mm)						
Mean (SD)	62 (± 8.5)	65 (± 8.5)	61 (± 8.5)	62 (± 8.7)	62 (± 8.5)	65 (± 8.5)
Biparietal diameter (mm)						
Mean (SD)	21 (± 2.7)	22 (± 2.8)	21 (± 2.8)	21 (± 2.8)	21 (± 2.8)	22 (± 2.8)
Head circumference (mm)						
Mean (SD)	77 (± 9.3)	80 (± 9.7)	76 (± 9.3)	76 (± 9.2)	77 (± 9.3)	79 (± 9.7)
Abnormal nose						
No	2056 (99.9%)	398 (97.3%)	244 (100%)	26 (92.9%)	2300 (99.9%)	424 (97.0%)
Yes	2 (0.1%)	11 (2.7%)	0 (0%)	2 (7.1%)	2 (0.1%)	13 (3.0%)
Fetal heart rate (times per minute)						
Mean (SD)	160 (± 9.2)	160 (± 10)	160 (± 8.8)	160 (± 8.8)	160 (± 9.1)	160 (± 9.9)
Nuchal translucency (mm)						
Mean (SD)	1.9 (± 1.1)	3.4 (± 1.0)	1.9 (± 1.1)	2.7 (± 0.83)	1.9 (± 1.1)	3.3 (± 1.0)
PAPP-A (MoM)						
Mean (SD)	0.74 (± 0.39)	0.53 (± 0.37)	0.73 (± 0.37)	0.66 (± NA)	0.74 (± 0.39)	0.53 (± 0.36)
β-hCG (MoM)						
Mean (SD)	1.3 (± 0.82)	2.0 (± 0.93)	1.5 (± 0.84)	3.7 (± NA)	1.3 (± 0.82)	2.0 (± 0.94)

Table 3.1 shows the summary of each indice in the trimester 1 dataset. We collected data of 2739 pregnant woman in trimester 1, of which 437 had a fetus with Down Syndrome (16%). The mean age of pregnant women with Down Syndrome was 32 years, which was 2 years higher on average than those without a Down fetus. Mean fetus's age was 90, fetus with Down had higher age. Fetuses with Down had the mean nuchal translucency thickness of 3.3, almost 2 times higher than the reference group of fetus without Down Syndrome. These fetuses also had lower mean PAPP-A

and higher β-hCG.

Table 3.2: Characteristics of participants in trimester 2

	Training		Testing		Overall	
	Not Down	Down	Not Down	Down	Not Down	Down
	(N=3983)	(N=626)	(N=742)	(N=60)	(N=4725)	(N=686)
Mother's age (years)						
Mean (SD)	31 (± 6.0)	33 (± 6.8)	32 (± 5.8)	33 (± 6.3)	31 (± 6.0)	33 (± 6.7)
Mother's age						
≥ 35	1169 (29 %)	273 (44 %)	251 (34 %)	28 (47 %)	1420 (30 %)	301 (44 %)
< 35	2772 (70 %)	353 (56 %)	488 (66 %)	32 (53 %)	3260 (69 %)	385 (56 %)
Fetus's age (weeks)						
Mean (SD)	130 (± 17)	120 (± 14)	130 (± 15)	120 (± 15)	130 (± 17)	120 (± 14)
History of having children with Down syndrome						
No	3955 (99 %)	619 (99 %)	683 (92 %)	58 (97 %)	4638 (98 %)	677 (99 %)
Yes	28 (1 %)	7 (1 %)	10 (1 %)	0 (0 %)	38 (1 %)	7 (1 %)
Biparietal diameter (mm)						
Mean (SD)	41 (± 8.5)	38 (± 7.2)	40 (± 7.1)	38 (± 8.2)	41 (± 8.3)	38 (± 7.3)
Head circumference (mm)						
Mean (SD)	150 (± 32)	140 (± 26)	150 (± 26)	140 (± 28)	150 (± 31)	140 (± 26)
Abnormal fetal nose						
No	3834 (96 %)	592 (95 %)	723 (97 %)	40 (67 %)	4557 (96 %)	632 (92 %)
Yes	149 (4 %)	34 (5 %)	19 (3 %)	20 (33 %)	168 (4 %)	54 (8 %)
Fetal heart rate (times per minute)						
Mean (SD)	150 (± 7.9)	150 (± 8.0)	150 (± 7.6)	150 (± 7.9)	150 (± 7.9)	150 (± 8.0)
Abnormal fetal heart						
No	3652 (92 %)	591 (94 %)	717 (97 %)	54 (90 %)	4369 (92 %)	645 (94 %)
Yes	331 (8 %)	35 (6 %)	17 (2 %)	5 (8 %)	348 (7 %)	40 (6 %)
Fetal femur length (mm)						
Mean (SD)	25 (± 7.7)	21 (± 6.1)	25 (± 6.5)	22 (± 6.7)	25 (± 7.5)	21 (± 6.2)
AFP (MoM)						
Mean (SD)	0.83 (± 0.26)	0.79 (± 0.25)	0.78 (± 0.27)	0.83 (± 0.46)	0.82 (± 0.27)	0.79 (± 0.26)
hCG (MoM)						
Mean (SD)	1.3 (± 0.81)	2.2 (± 0.91)	1.5 (± 0.83)	2.8 (± 1.6)	1.4 (± 0.82)	2.2 (± 0.93)
uE3 (MoM)						
Mean (SD)	0.92 (± 0.37)	0.71 (± 0.25)	0.73 (± 0.33)	0.57 (± 0.22)	0.88 (± 0.37)	0.71 (± 0.25)

In trimester 2, we collected data of 5411 women, of which 686 pregnant women had a fetus with Down syndrome (12.7%). Women with a Down fetus were, on average, 2 years older than those without Down Syndrome. Fetus's age was 10 weeks younger in the Down group. Fetus with Down Syndrome also had lower biparietal diameter, head circumference and femur length than those who did not

have Down Syndrome. In the 3 triple test indices, Down fetus had lower AFP and uE3 values, with higher hCG as compared with the reference group.

3.1.3 Features of each dataset

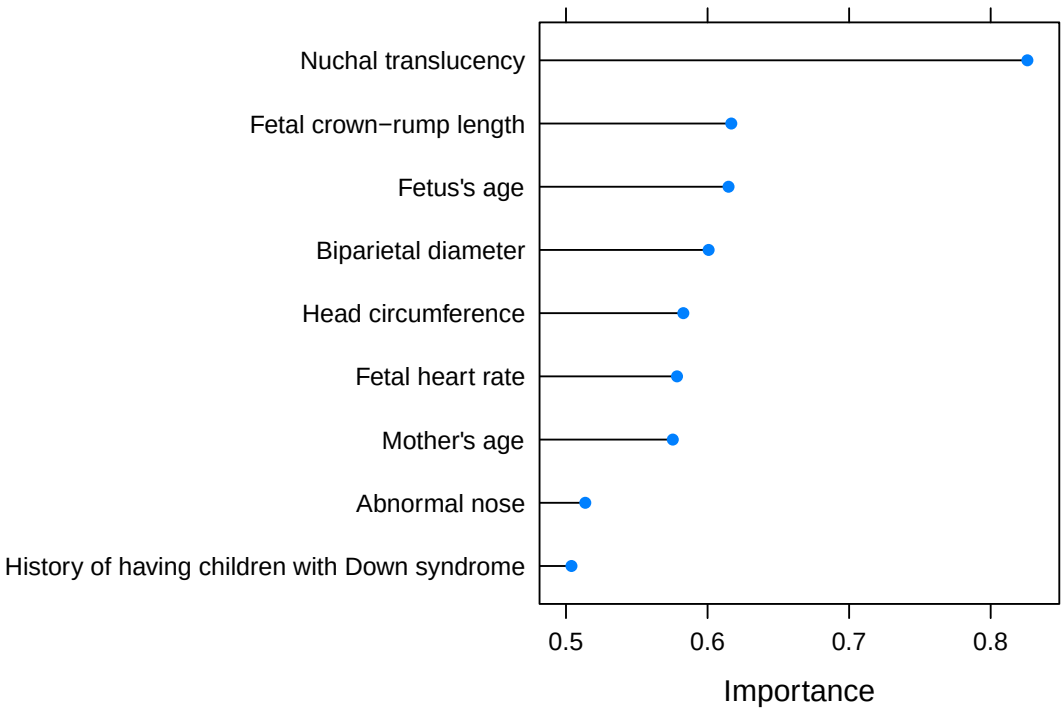

Figure 3.3: Ranking variables importance in trimester 1 ultrasound data

In the trimester 1 ultrasound dataset, the importance of 9 variables in the prediction of Down Syndrome is shown in figure 3.3. The feature on top was the one that contributed the most to the prediction of Down Syndrome and the one at the bottom contributed the least in this dataset. Therefore, nuchal translucency was the most important variable and the mother's history of having children with Down Syndrome had the lowest importance.

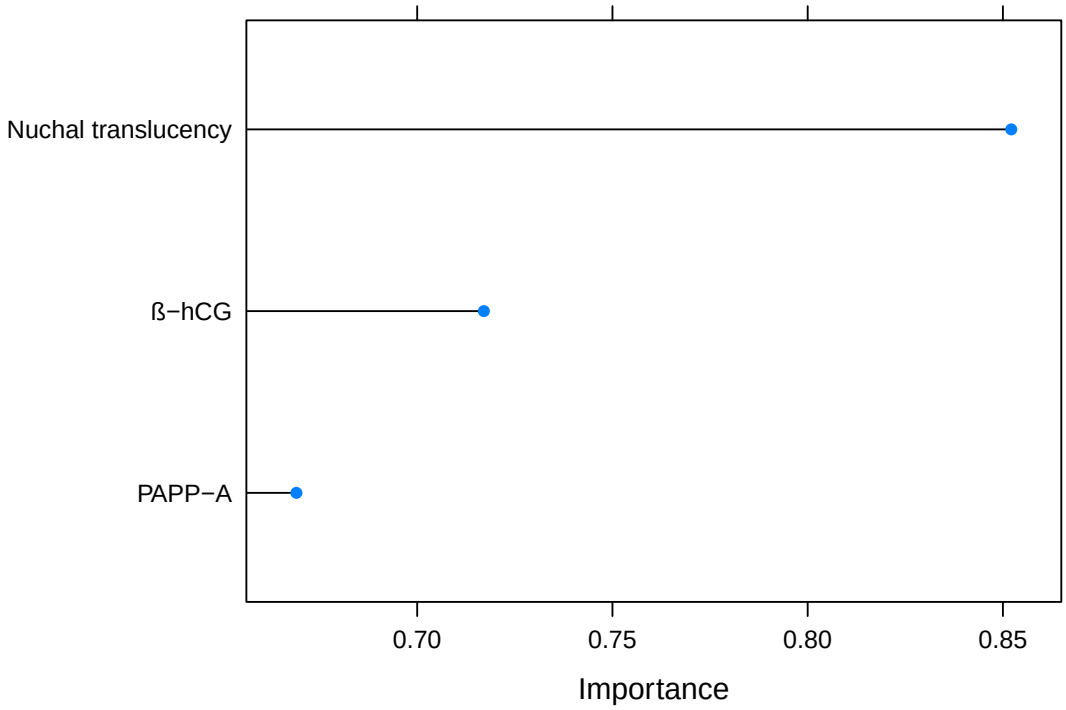

Figure 3.4: Ranking variables importance in trimester 1 biochemical data

Nuchal translucency was the most important variable in the trimester 1 biochemical dataset as shown in figure 3.4, followed by β-hCG and PAPP-A.

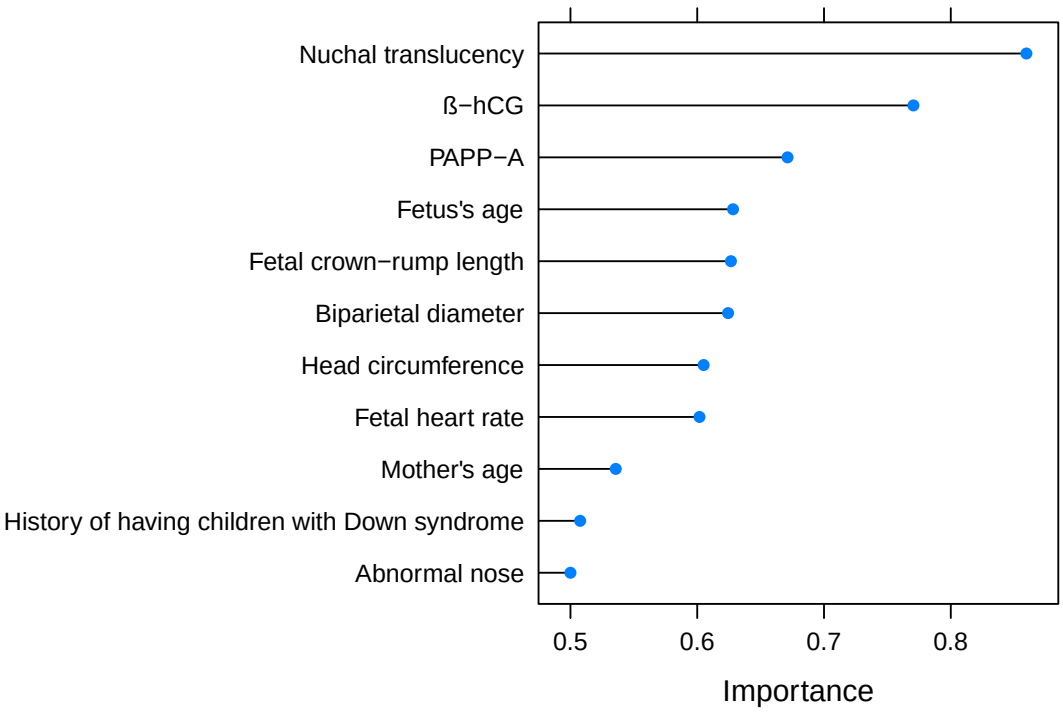

Figure 3.5: Ranking variables importance in trimester 1 ultrasound and biochemical data

In the dataset that had both ultrasound and biochemical data, nuchal translucency and biochemical variables continued to be most important features in the whole dataset.

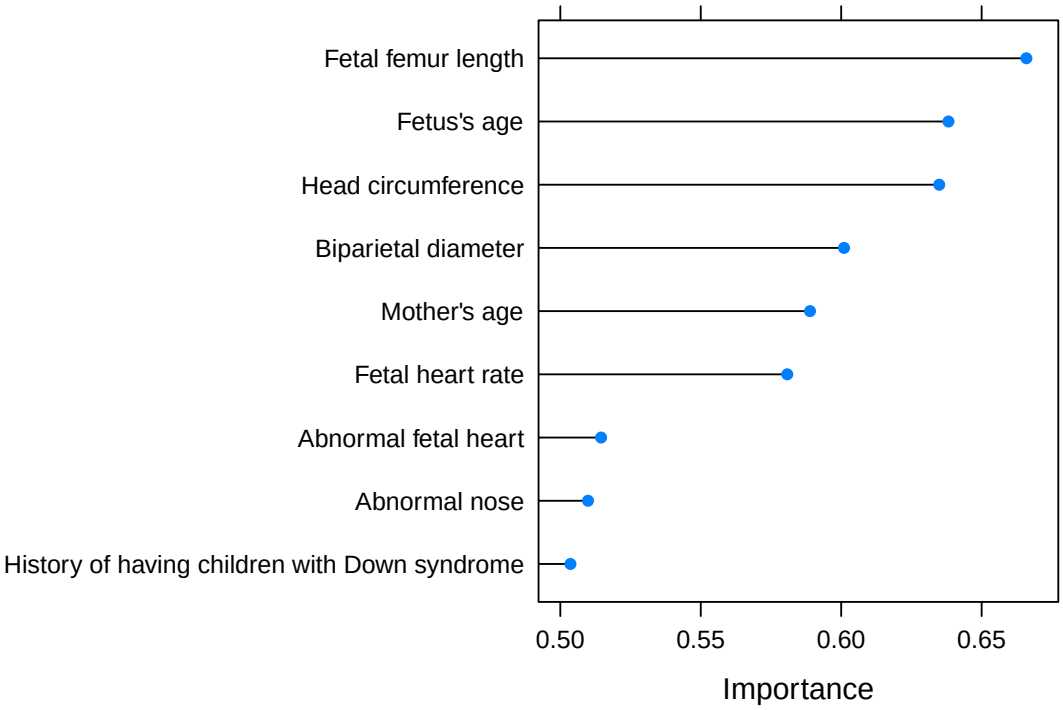

Figure 3.6: Ranking variables importance in trimester 2 ultrasound data

In trimester 2, nuchal translucency was no longer used in the prediction models. Fetal femur length was the most important feature in the trimester 2 ultrasound data, followed by fetus age, head circumference and the mother's history of having children with Down Syndrome continued to have the lowest importance score.

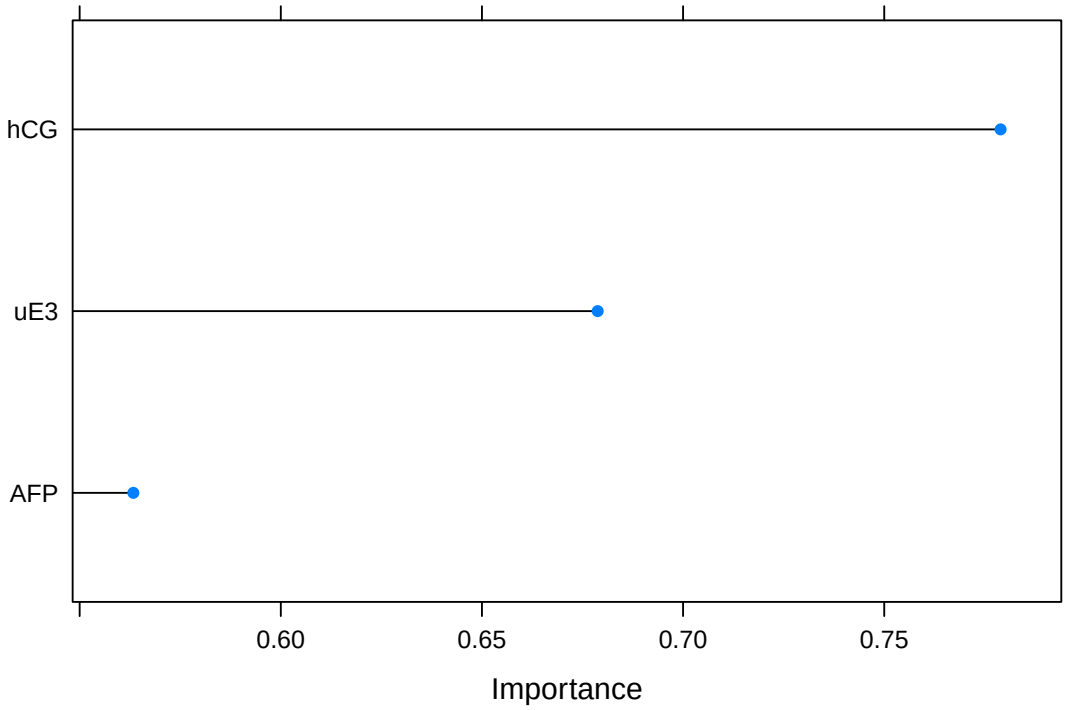

Figure 3.7: Ranking variables importance in trimester 2 biochemical data

As shown in figure 3.7, hCG had the highest importance score in the trimester 2 biochemical dataset.

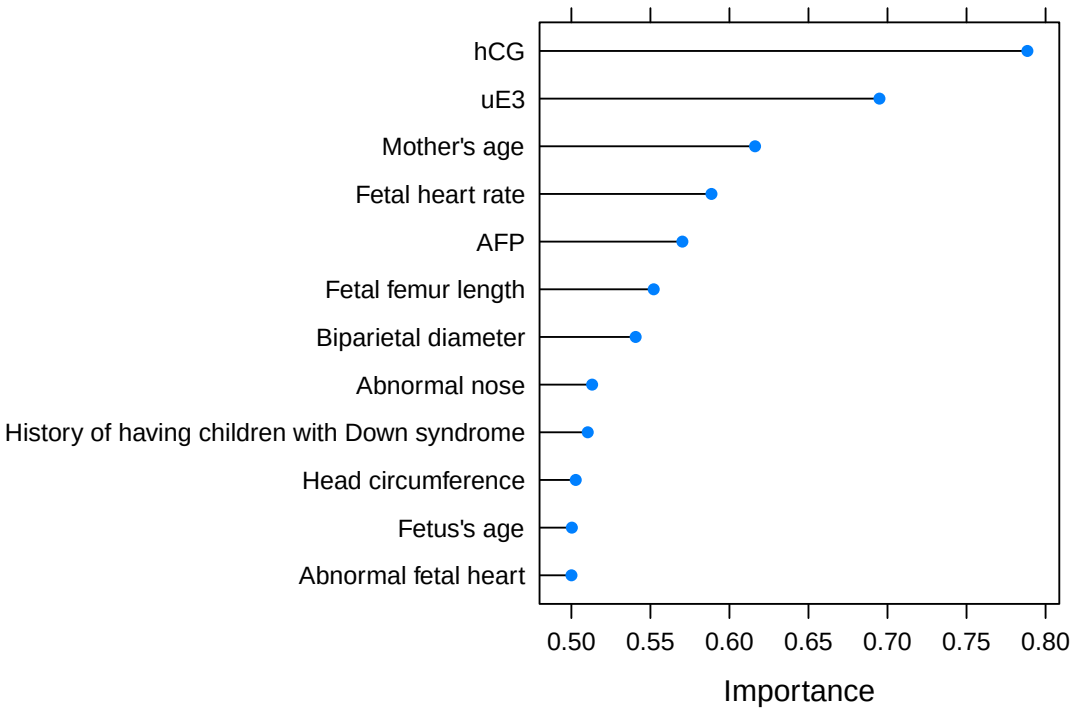

Figure 3.8: Ranking variables importance in trimester 2 ultrasound and biochemical data

In the combined dataset, trimester 2 exhibited a similar pattern to trimester 1, where biochemical values remained the most influential factor in predicting Down Syndrome. Interestingly, fetal age, initially ranked as the second most important variable, dropped to become the second least significant variable in the prediction models.

3.1.4 Model performance of each machine learning model in each dataset

Table 3.3: Model performance in training process on trimester 1 data

Model	ROC	Sens	Spec	Accuracy
Ultrasound				
k-nearest neighbor	0.90	0.74	0.92	0.82
Support Vector Machine	0.85	0.78	0.83	0.80
Random Forest	0.82	1.00	0.02	0.85
XGBoost	0.83	0.93	0.33	0.84
Biochemical				
k-nearest neighbor	0.96	0.81	0.95	0.88
Support Vector Machine	0.92	0.77	0.77	0.77
Random Forest	0.88	0.98	0.50	0.94
XGBoost	0.92	0.99	0.41	0.94
Both ultrasound and biochemical				
k-nearest neighbor	0.99	0.82	0.99	0.91
Support Vector Machine	0.94	0.87	0.89	0.88
Random Forest	0.89	1.00	0.30	0.95
XGBoost	0.89	0.99	0.45	0.95

3.1.4.1 Trimester 1 The performance of the models in trimester 1 during the training phase is presented in 3.3. The highest accuracy achieved was 95% by RF and XGBoost when using the combined dataset. These two models also achieved the highest accuracy of 94% with the biochemical dataset. In the ultrasound dataset, the RF model exhibited 85% accuracy, which was 1% higher than the accuracy of the XGBoost model.

Table 3.4: Model performance in validating process on trimester 1 data

Model	threshold	specificity	sensitivity	ppv	npv	accuracy	tp	tn	fp	fn
Ultrasound										
k-nearest neighbor	0.44	0.81	0.80	0.44	0.96	0.81	51	284	66	13
Support Vector Machine	0.68	0.74	0.89	0.39	0.97	0.76	57	259	91	7
Random Forest	0.91	0.84	0.69	0.44	0.94	0.82	44	295	55	20
XGBoost	0.83	0.87	0.75	0.51	0.95	0.85	48	303	47	16
Biochemical										
k-nearest neighbor	0.28	0.91	0.63	0.38	0.97	0.88	12	194	20	7
Support Vector Machine	0.50	0.91	0.79	0.44	0.98	0.90	15	195	19	4
Random Forest	0.52	0.98	0.58	0.69	0.96	0.94	11	209	5	8
XGBoost	0.63	0.97	0.63	0.63	0.97	0.94	12	207	7	7
Both ultrasound and biochemical										
k-nearest neighbor	0.29	0.91	0.80	0.43	0.98	0.91	12	172	16	3
Support Vector Machine	0.54	0.85	0.93	0.33	0.99	0.85	14	159	29	1
Random Forest	0.91	0.84	0.80	0.29	0.98	0.84	12	158	30	3
XGBoost	0.91	0.94	0.73	0.50	0.98	0.93	11	177	11	4

In the validation process, the XGBoost model achieved the highest accuracy in all three datasets: 85% in ultrasound, 94% in biochemical, and 93% in the combined dataset. The RF model secured the second-best accuracy in the ultrasound and biochemical datasets.

Table 3.5: Model performance in training process on trimester 2 data

Model	ROC	Sens	Spec	Accuracy
Ultrasound				
k-nearest neighbor	0.80	0.63	0.80	0.72
Support Vector Machine	0.76	0.67	0.73	0.70
Random Forest	0.69	0.97	0.12	0.85
XGBoost	0.71	0.97	0.09	0.85
Biochemical				
k-nearest neighbor	0.85	0.65	0.89	0.77
Support Vector Machine	0.83	0.73	0.74	0.73
Random Forest	0.81	0.97	0.28	0.86
XGBoost	0.80	0.96	0.15	0.84
Both ultrasound and biochemical				
k-nearest neighbor	0.86	0.56	0.93	0.74
Support Vector Machine	0.86	0.78	0.81	0.79
Random Forest	0.82	0.97	0.32	0.87
XGBoost	0.83	0.98	0.23	0.86

3.1.4.2 Trimester 2 Both RF and XGBoost models emerged as the top performers during the training process of trimester 2 (table 3.5). All RF and XGBoost models demonstrated accuracy ranging between 84% and 87%, while the kNN and SVM models achieved accuracies ranging from 70% to 80%.

Table 3.6: Model performance in validating process on trimester 2 data

Model	threshold	specificity	sensitivity	ppv	npv	accuracy	tp	tn	fp	fn
Ultrasound										
k-nearest neighbor	0.46	0.70	0.57	0.24	0.91	0.68	63	477	201	48
Support Vector Machine	0.55	0.59	0.75	0.23	0.93	0.61	83	401	277	28
Random Forest	0.77	0.79	0.46	0.27	0.90	0.75	51	538	140	60
XGBoost	0.85	0.74	0.54	0.25	0.91	0.71	60	499	179	51
Biochemical										
k-nearest neighbor	0.40	0.79	0.65	0.35	0.93	0.77	24	167	45	13
Support Vector Machine	0.57	0.69	0.76	0.30	0.94	0.70	28	146	66	9
Random Forest	0.81	0.79	0.59	0.33	0.92	0.76	22	167	45	15
XGBoost	0.67	0.86	0.62	0.43	0.93	0.82	23	182	30	14
Both ultrasound and biochemical										
k-nearest neighbor	0.34	0.87	0.64	0.47	0.93	0.83	21	157	24	12
Support Vector Machine	0.26	0.94	0.64	0.66	0.93	0.89	21	170	11	12
Random Forest	0.70	0.88	0.64	0.49	0.93	0.84	21	159	22	12
XGBoost	0.73	0.90	0.70	0.56	0.94	0.87	23	163	18	10

The highest accuracy in the trimester 2 validation process was achieved by the SVM model with a combined dataset, reaching 89%. This was followed by XGBoost in the same dataset. In the biochemical dataset, XGBoost achieved the highest accuracy, while RF performed the best with the ultrasound dataset.

3.2 Performance of each machine learning model on the test set

3.2.1 Trimester 1

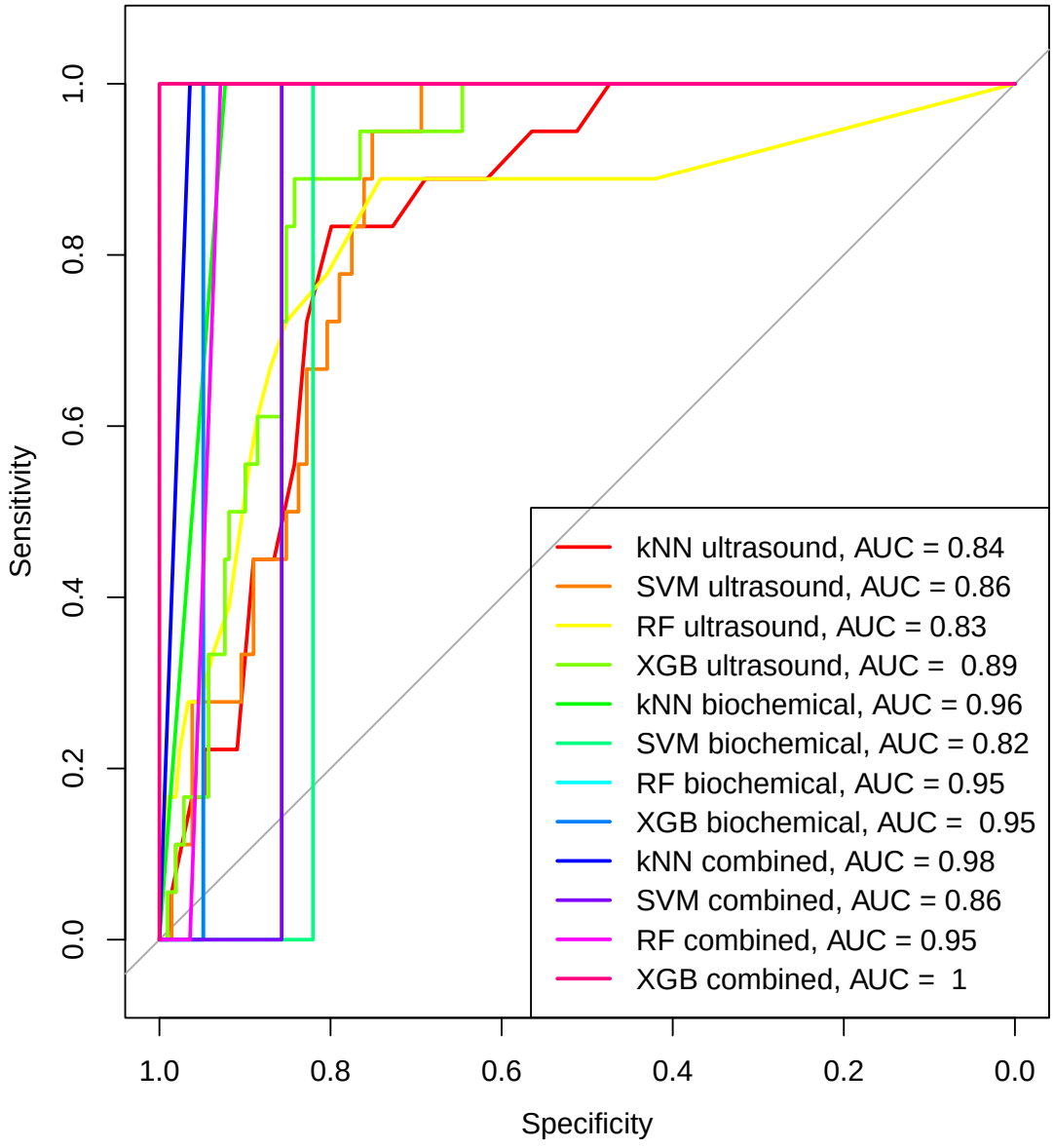

Figure 3.9: Comparison of models on trimester 1 testing data through ROC curve

Figure 3.9 illustrates the ROC curves of all 12 trimester 1 machine learning models for their respective test sets. All 12 models exhibited AUC values higher than 0.8. The XGBoost model built with the combined dataset achieved the highest AUC,

followed by kNN in the same dataset. XGBoost also attained the highest AUC in the ultrasound dataset. In the biochemical dataset, kNN achieved the best AUC of 0.96, followed by RF and XGBoost, both with AUC values of 0.95.

Table 3.7: Models performance in testing process on trimester 1 data

Model	threshold	specificity	sensitivity	ppv	npv	accuracy	tp	tn	fp	fn
Ultrasound										
k-nearest neighbor	0.40	0.80	0.83	0.26	0.98	0.80	15	167	42	3
Support Vector Machine	0.65	0.75	0.94	0.25	0.99	0.77	17	157	52	1
Random Forest	0.95	0.74	0.89	0.23	0.99	0.75	16	155	54	2
XGBoost	0.82	0.84	0.89	0.33	0.99	0.85	16	176	33	2
Biochemical										
k-nearest neighbor	0.06	0.92	1.00	0.25	1.00	0.92	1	36	3	0
Support Vector Machine	0.50	0.82	1.00	0.12	1.00	0.82	1	32	7	0
Random Forest	0.41	0.95	1.00	0.33	1.00	0.95	1	37	2	0
XGBoost	0.60	0.95	1.00	0.33	1.00	0.95	1	37	2	0
Both ultrasound and biochemical										
k-nearest neighbor	0.04	0.96	1.00	0.50	1.00	0.97	1	27	1	0
Support Vector Machine	0.13	0.86	1.00	0.20	1.00	0.86	1	24	4	0
Random Forest	0.58	0.93	1.00	0.33	1.00	0.93	1	26	2	0
XGBoost	0.41	1.00	1.00	1.00	1.00	1.00	1	28	0	0

The performance of each of the trimester 1 models during the testing phase is presented in Table 3.7. All machine learning models that were trained on datasets containing biochemical values correctly identified the single case of Down Syndrome. Similarly, in terms of AUC, the XGBoost model trained on the combined datasets achieved the highest overall accuracy, achieving a perfect score of 1 by correctly classifying all 29 fetuses. This was followed by the kNN model trained on the same dataset. As observed in the training and validation phase, the RF and XGBoost models achieved the highest accuracy in the biochemical dataset, with XGBoost emerging as the superior model in the ultrasound dataset.

3.2.2 Trimester 2

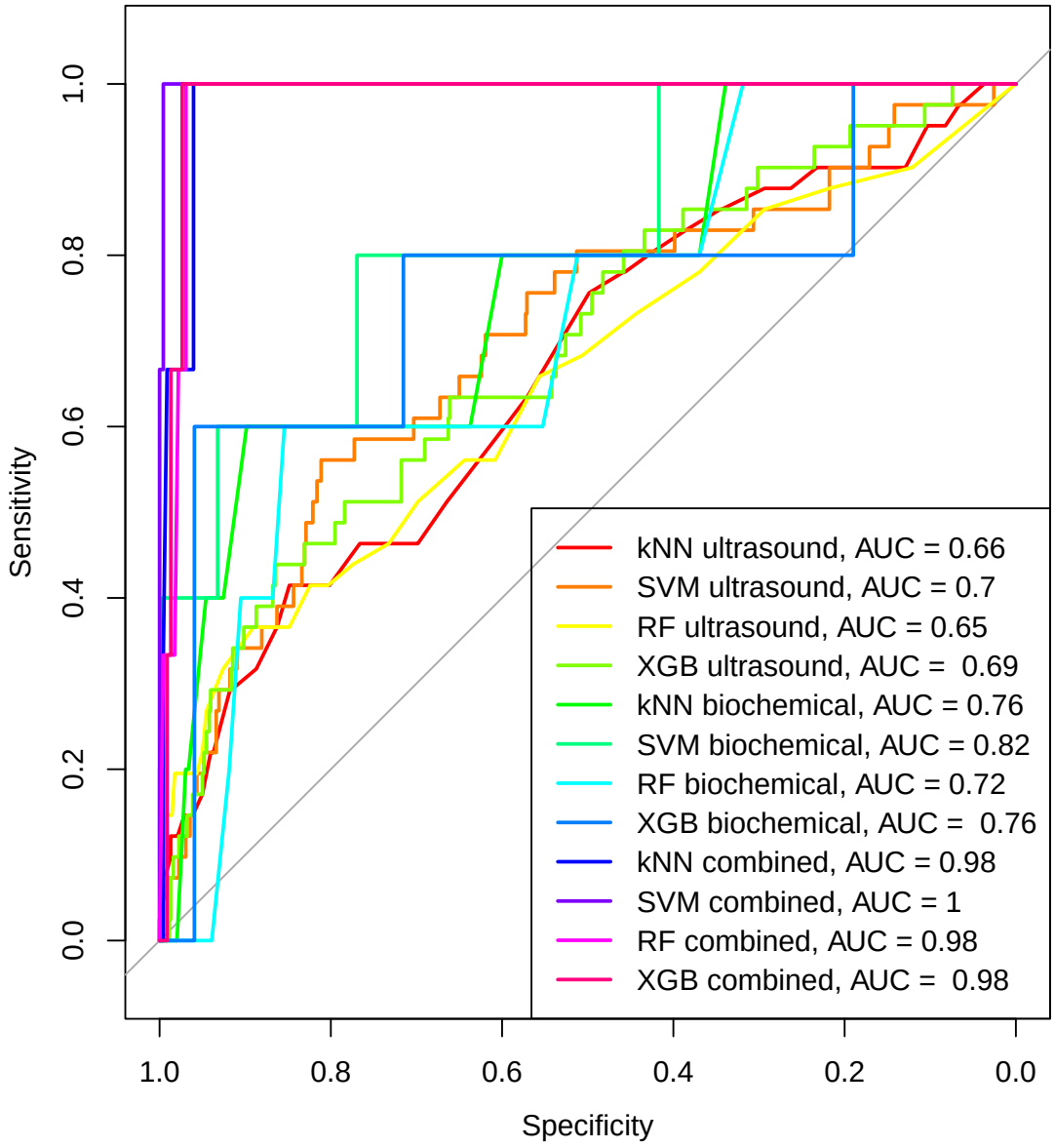

Figure 3.10: Comparison of models on trimester 2 testing data through ROC curve

Using the trimester 2 data, all four models constructed using the combined dataset attained the highest AUC scores. There was no clear different in ROC curve between models built using ultrasound and biochemical dataset. The top-performing model in this trimester was SVM, boasting an AUC of 1. The other three models all

achieved an AUC score of 0.98.

Table 3.8: Models performance in testing process on trimester 2 data

Model	threshold	specificity	sensitivity	ppv	npv	accuracy	tp	tn	fp	fn
Ultrasound										
k-nearest neighbor	0.35	0.85	0.41	0.15	0.96	0.82	17	526	94	24
Support Vector Machine	0.39	0.81	0.56	0.16	0.97	0.80	23	503	117	18
Random Forest	0.67	0.89	0.37	0.18	0.96	0.86	15	552	68	26
XGBoost	0.76	0.86	0.44	0.18	0.96	0.84	18	536	84	23
Biochemical										
k-nearest neighbor	0.18	0.90	0.60	0.09	0.99	0.89	3	265	30	2
Support Vector Machine	0.36	0.77	0.80	0.06	1.00	0.77	4	227	68	1
Random Forest	0.65	0.85	0.60	0.07	0.99	0.85	3	252	43	2
XGBoost	0.52	0.96	0.60	0.20	0.99	0.95	3	283	12	2
Both ultrasound and biochemical										
k-nearest neighbor	0.25	0.96	1.00	0.25	1.00	0.96	3	218	9	0
Support Vector Machine	0.02	1.00	1.00	0.75	1.00	1.00	3	226	1	0
Random Forest	0.42	0.97	1.00	0.30	1.00	0.97	3	220	7	0
XGBoost	0.51	0.97	1.00	0.33	1.00	0.97	3	221	6	0

Table 3.8 provides a detailed breakdown of the results for all models during the testing phase of trimester 2 data. Models generated from the combined datasets accurately detected all three cases of Down Syndrome. SVM recorded the highest AUC and accuracy, correctly classifying 226 out of 227 negative cases of Down Syndrome. Both RF and XGBoost achieved a 97% accuracy rate with the same dataset, followed by kNN. In the biochemical dataset, XGBoost outperformed other models with an accuracy of 95%. However, in the ultrasound dataset, RF surpassed XGBoost with an 86% accuracy rate.

4 Chapter 4: Discussion

In this study, we wanted to build and find the best and most suitable machine learning model for each type of situations in practice. There would be medical facilities in low level health care system that can only do the ultrasound test, only the biochemical test or both of the test. By making models for every type of situations, we can expand the scope of this screening program and let more pregnant women be screened for Down Syndrome, thus lowering the prevalence of Down Syndrome cases in the community. Machine learning models was a cheap yet effective and could be applied in low level health care setting where didn't have medical expert in prenatal screening.

4.1 Dataset characteristics

We chose to build models according to 3 sets of ultrasound, biochemistry and combined data based on the fact that ultrasound is one of the most popular testing methods in medical facilities and biochemical testing is the most frequently used method of prenatal screening.

4.1.1 Included variables

In the ultrasound dataset, the selection of an index as the input variable for the machine learning model was based on the number of missing that the variable had. The selected variables are the ones with the least number of missing in the ultrasound indexes. Variables that had more missing show little value in prenatal screening in practice. With what variables a machine learning model is built, it needs to have those variables as input when put into practice to be able to produce results. The fact that there are many missing values in the dataset to build the models also partly reflects this situation in reality. So even if these variables are added to the machine learning model, it will be difficult to apply them in practice.

With the biochemical dataset, the 3 input variables in each trimester are routine indices in the double test and triple test for trimester 1 and trimester 2 respectively. Due to the fact that one health care facility could have many different types of testing machines with different scales and standards, not to mention that if these machine learning models were applied in practice with different medical facilities, the diversity could also be multiplied many times over. To solve this problem, we need to have a unit common to all analyzers, which is the multiple of median (MoM). A multiple of the median (MoM) is a measure of how far an individual test result deviates from the median .

An MoM for a test result for a patient can be determined by the following:

$$MoM(Patient) = \frac{Result(Patient)}{Median(PatientPopulation)}$$

4.1.2 Training, validating and testing dataset

In order to be included in the 6 sub-dataset, each case must have values of all the required variables, which means the combined dataset must have less observations than ultrasound or biochemical dataset.

The training dataset was used to build machine learning models. The k cross validation method means that the training dataset were divided into k equal parts, of which k-1 parts were used to build the model and 1 part was used for testing. In this topic, we used 10 cross validation method, that is, 9/10 parts were used to build models. The performance in training process showed in the results section were the results of testing built models on 1/10 the number of observations of these training dataset. This process was done to fine-tune model parameters and assess model performance during training phase.

The next step involves validating these models. Validation is the process of testing models using a separate validation set that shares the same characteristics as the training set. This validation process was carried out to assess machine learning

models and refine them before practical application. In this context, “refine” means evaluating whether the model is overfitting and adjusting its parameters accordingly.

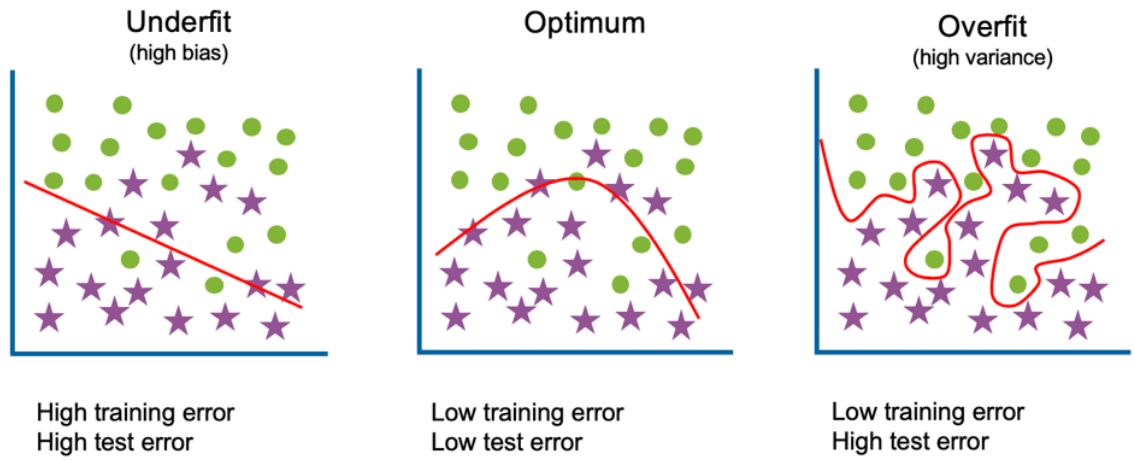

Figure 4.1: Demonstration of overfitting

Figure 4.1 provides an illustrative example of overfitting. Overfitting occurs when a model fits the data too closely, attempting to capture every single data point it encounters. This can lead to the model capturing noise and inaccuracies from the dataset, ultimately degrading its performance. Therefore, in this study, the models were fine-tuned to mitigate overfitting. This is evident from the minimal difference in accuracy between the training results and the validation results, indicating that the models were effectively adjusted to maintain their performance.

The test set, on the other hand, contained observations that were not used in the training and validating process. It was a completely new, unseen dataset that was collected from another hospital. In this cases, test data was collected from 3 different hospitals which were Hanoi Medical University hospital, Hanoi Obstetrics and Gynecology Hospital and the National Hospital of Obstetrics and Gynecology. The purpose of test set was to see how well our models fit with real life unseen data.

Figure 4.2 demonstrates an example of test data by showing the diversity of features in different type of cups[27]. There are 10 type of cups in figure. Although

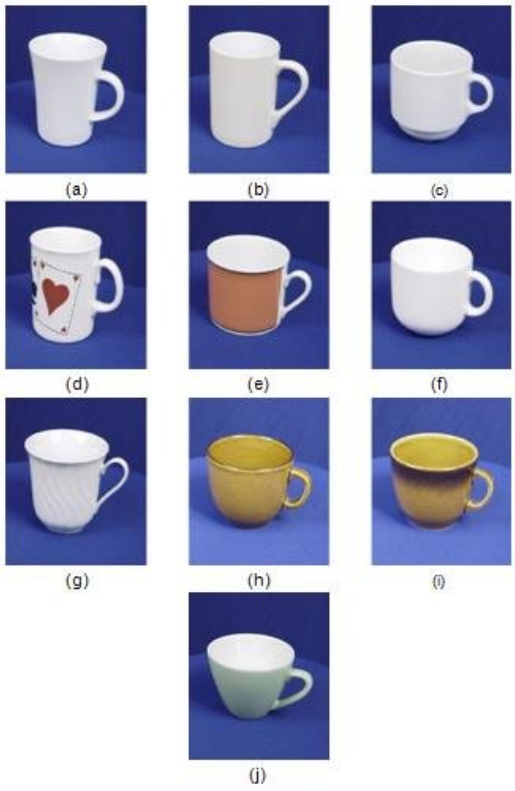

Figure 4.2: Different type of cups

they are all called cups, they differ in length, body width, bottom width, color, handle size, etc. We can take for example the picture of cups from (a) to (g) as the training and validating dataset and (h) to (j) as the test dataset since the cups from (h) to (j) has characteristics that are different from cups (a) through (g). This test is intended to evaluate on test data how well machine learning models generalize a definition of a cup from the training data.

4.1.3 Study population characteristics

Pregnant women with Down Syndrome had higher mean age by 2 years in both trimester. The same pattern happened when we divide age by 2 groups, equal or more than 35 and less than 35. As we mentioned earlier, current literature shows the relationship between age and the risk of pregnant a baby with Down. This can be explained by aging in the reproductive organs leading to uneven chromosome division

during egg production in women[28].

History of having children with Down Syndrome didn't contributed much in the prediction of Down in our dataset since there were only around 1% of maternal reported to had Down children in the past in both trimester. There is always a small percentage of errors occurring during the egg division process due to the influence of external factors at that time instead of a failure systematically leading to a deviation in the number of chromosomes, thus did not produce a different in our result.

It's recommended to do regular screening check up during 11 to 13 weeks 6 days of pregnancy for double test and 15 weeks to 20 weeks 6 days for triple test. Pregnant women in our study had a later check-up time than recommended, but most were still within the time period.

4.2 Feature importance

4.2.1 Trimester 1

Nuchal translucency was the most important variable in predicting Down Syndrome in all dataset containing this variable in trimester 1. This result was relevant with current literature as the importance of predicting Down syndrome using nuchal translucency had been demonstrated in various studies. Anna Locatelli in 2000 concluded that the use of the difference between observed and expected nuchal fold thicknesses to determine likelihood ratios allows the calculation of individual posterior probabilities of Down syndrome[29]. In 1995, Dr B.Brambati confirmed the potential application of the measurement of nuchal translucency thickness for fetal aneuploidy screening before the end of the first trimester. Measurement of nuchal translucency in predicting trisomies was now a routine procedure in predicting Down Syndrome[30].

Our findings indicated that the mother's age and history of having Down Syndrome were not as significant as other ultrasound variables in predicting Down Syndrome, while current literature supports the idea that women over 35 years old are at a

higher risk of having a pregnancy affected by Down Syndrome. This can be explained by the fact that the results of these ultrasound and biochemical tests are directly influenced by the fetal Down Syndrome condition. When compared to these factors, an indirectly influenced factor like the mother's age would likely have limited ability to predict the Down Syndrome condition equivalently. This finding was support by another study that measure feature importance on mother's age and 6 others biochemical values which produced similar result[31].

4.2.2 Trimester 2

In trimester 2, femur length was the most important variables in the ultrasound dataset. This finding was also proven by previous literature. Femur length together had been shown to be two important indicators in predicting Down in the second trimester[32–35]. Biparietal diameter was also an essential index as biparietal diameter/femur length ratio was found to decrease with gestational age in the normal population and was consistently elevated in the Down syndrome population throughout the second trimester.

In a study conducted by Kevin Spencer in 2005 to assess the impact and value of AFP and hCG levels in screening for trisomy 21 and trisomy 18 among 67,904 pregnant women, the author concluded that the hCG index can achieve high detection rates over an extended period of time[36]. In another study by Vivienne L. Souter in 2002 involving 72 pregnant women with Down Syndrome, the triple test indices were combined with ultrasound to explore the correlation when combining these two sets of indices in screening[37]. The results demonstrated that the hCG index, when combined with nuchal translucency measurement, showed the highest correlation. These studies align with our findings in assessing the importance of hCG in the trimester 2 biochemical dataset.

In both trimester, both combined dataset, biochemical variables were shown to have higher impact on the prediction of Down rather than ultrasound indices. This

explains the specific use of these biochemical indices in screening for trisomies in the current universal screening program.

4.3 Model performance

4.3.1 Hyper-parameters used

We used 10 fold validation and validating set on all models to search for the best hyper-parameters. During this process, we used grid search strategy with the goal was to minimize the difference in accuracy between testing and validating result while keeping both of them at the highest value in order to reduce over fitting while maintaining their performance. The hyper-parameters that we searched for each model is shown in detail below:

1. kNN: k as the number of nearest neighbors to consider when making a prediction or classification
2. SVM: C as the parameter that controls the penalty for margin violations, which are data points that fall on the wrong side of the decision boundary or within the margin
3. RF:

- mtry as a hyperparameter that controls the number of features (variables or predictors) randomly selected as candidates for splitting at each tree node during the tree-growing process
- ntree as the number of trees in the model
- max_depth (depth) as the maximum depth of tree

4. XGBoost:

- nrounds determines how many decision trees (weak learners) are sequentially added to the ensemble

- min_child_weight (mcw) controls the minimum sum of instance weight (hessian) needed in a child
- subsample (ssp) controls the fraction of training data to be randomly sampled during each boosting round
- colsample_bytree (csb) determines the fraction of variables to be randomly sampled when building each tree
- max_depth (xgbdepth) sets the maximum depth of each tree in the ensemble
- eta (learning rate) controls the step size or learning rate used in the gradient boosting process

Table 4.1: Detailed hyper-parameters of each model

	kNN	SVM	RF			XGBoost					
	k	C	mtry	ntree	depth	nrounds	mcw	ssp	csb	xgbdepth	eta
Trimester 1											
T1 ultrasound	26	0.1000	1	50	3	200	2	0.6	0.6	5	0.10
T1 biochemical	9	0.0001	3	50	3	100	5	1.0	1.0	3	0.10
T1 combined	12	0.5000	2	200	3	200	1	0.6	0.6	5	0.10
Trimester 2											
T2 ultrasound	36	0.0200	5	50	3	200	5	1.0	0.6	7	0.10
T2 biochemical	59	0.4000	1	50	3	100	1	0.8	1.0	5	0.01
T2 combined	169	0.0100	7	200	3	50	1	1.0	0.6	3	0.10

The detail hyper-parameters of each model is shown in 4.1. With the most hyper-parameters, XGBoost was the model that required the most time and computing power with huge number of parameters combination. Yet it is the best model with good adaptability and flexibility. Thus it had the highest accuracy in most of our dataset in comparison with other models. XGBoost builds a series of multiple models, each model learns to correct the errors of the previous model (the data that the previous model predicted incorrectly) to be better than the previous model. The main method is to increase the weight of incorrectly predicted data, keeping the weight of correctly predicted data unchanged to train the next model. This model had already been proven as one of the best model in classification problem and it had been the dominant model

in many machine learning competitions[38–40].

4.3.2 Model accuracy

4.3.2.1 Better result by adding more variables The models constructed using the combined dataset consistently outperformed their counterparts in both trimesters. In trimester 1, while the model built on the combined dataset slightly outperformed the model using only the biochemical dataset, both significantly outshone the model relying solely on ultrasound data. This suggests that the integration of biochemical data substantially enhances the accuracy of Down Syndrome screening during the early stages of pregnancy.

The superiority of the combined dataset model became even more evident in trimester 2, where it far outperformed the other two models. This reinforces the notion that a multi-faceted approach, encompassing both ultrasound and biochemical data, provides the most reliable screening results as pregnancy progresses.

So there was a clear result that models built from the combined data set produced more accurate results than models built from the other two data sets. This showed that combining general, ultrasound and biochemical information indicators increases the accuracy of the machine learning model for prenatal screening for Down syndrome. The general theory of machine learning models mentioned increasing the efficiency of the model by adding more variables[41]. However, when we attempted to include both the mother's age and fetus's age in the model constructed with the biochemical dataset for both trimesters, the outcomes did not exhibit the same level of success, although they showed a great importance through feature calculation. Therefore adding more features into the model is not always a good practice.

4.3.2.2 Better result by pre-processing data The practice of pre-processing is also recommended when building machine learning models. We performed three pre-processing method which were scaling, centering and oversampling with kNN and

SVM model, leaving out RF and XGBoost. This final choice was given after trying to pre-process these two models which yield a not positive result in both training and validating phase. Each of the three pre-processing method serves a specific purpose:

- Scaling is used to normalize the range of variables in the dataset. This can be done by normalization(all values are shifted and rescaled so that they end up ranging from 0 to 1) or standardization(all value is centered with a mean of 0 and a standard deviation of 1).
- Centering subtract all the value by its variable's mean to adjust the distribution of a feature so that it has mean as the specific center point.

By applying scaling and centering, data points would be closer to each others, eliminate the influence of outliers on classification. That's why it's required for kNN and SVM, the two models that classify based on distances between data points.

The imbalance in data can lead to bias toward the majority class, leading the model to classify a case to be in the majority class more frequently. Model built using unbalanced dataset may struggle to generalize to new, unseen data, especially for the minority class because they have not been exposed to enough examples of the minority class during training. We had also tried training kNN and SVM model on unbalanced dataset, but the result was not as we expected. In the case of RF and XGBoost model, these two models have high robustness due to their ensemble nature and internal mechanisms which can handle imbalanced datasets well without requiring oversampling. This had been tried and the unbalanced model gave better results. The huge range of difference between sensitivity and specificity of all RF and XGBoost models during training phase was caused by this imbalance in training data. Despite that, results during validating and testing phase had blurred these differences, proving that they was not influence by the imbalance.

Not only did not pre-process the RF and XGBoost models yielded better results, but it also exposed the two models to real, unprocessed data, which the pre-

dictive models would continue to do in real practice. A study conducted in 2022 by Ahmad S. Tarawneh et al., involving 70 oversampling methods and 9 distinct real-world dataset, revealed that machine learning models trained on artificially generated data, especially when minority values were over-represented, tended to perform better in testing than in practical, real-world scenarios[42]. This is crucial in the case of Down Syndrome since the prevalence of Down Syndrome cases is small in the community. Thus, SVM model built on combined dataset shown highest accuracy in trimester 2 need to be further assess before deploying in real life practice.

4.3.2.3 The problem with small number of test cases In our dataset, here's 1 case of Down Syndrome in trimester 1 biochemical and combined test set, 5 cases in trimester 2 biochemical test set and 3 in trimester 2 combined test set which may not be sufficient amount of positive cases in some extends. However, the ability of machine learning models can also be measure by specificity which is the ability to correctly classify all of the negative cases beside sensitivity which is the ability to correctly classify all of the positive cases as well. The problem with a small number of positive cases in test set lead to a huge decrease in sensitivity if a new positive cases is not detected by the model. If the number of cases used to test model continue to increase, we expected the accuracy to be close to validation phase result due to the indifferent in the distribution of variables between validate set and test set.

4.3.2.4 Comparison with current Down screening method In screening for Down Syndrome, sensitivity is as importance as specificity. High sensitivity meaning that more cases of Down can be detected through screening when high specificity ensures that those cases are exactly Down cases. In the case of high sensitivity while having low specificity, more more pregnant women will be classify as Down Syndrome, and they need to do amniocentesis test to confirm the condition even though they do not have it. This not only causes unnecessary psychological and financial impacts for pregnant women and their families, but more importantly, it can

also affect the fetus and the mother's reproductive health. Although amniocentesis is a confirmatory test and the gold standard, it is an invasive test with complications such as miscarriage, rupture of membranes, and infection[43,44]. Thus, minimizing the number of pregnant women undergo this test through ensuring high specificity is a requirement of Down screening methods. Based on these evidences, we tried to choose the cut-off point with the highest accuracy to keep the balance between sensitivity and specificity.

A systematic review conducted on 56 studies, 204,759 pregnancies of which 2113 had pregnancies affected by Down in trimester 1 showed the sensitivity of the best Double test to be around 68% while the best specificity would be around 95%, which was outshined by our best machine learning model, even with the validating result, not to mention testing result. Another study conducted by the same group of authors on 59 studies involving 341,261 pregnancies with 1,994 cases of Down Syndrome in trimester 2 demonstrated similar result of the best model, around 60-70% sensitivity with 95% specificity, that also bested by our built models. Therefore, we concluded that our machine learning models had higher sensitivity and specificity than the current universal Down screening method.

In comparison with NIPT, despite that our best model, XGBoost on combined dataset in trimester 1 and SVM on combined dataset in trimester 2 showed the near 100% sensitivity and 100% specificity performance which is as high as NIPT, further testing need to be conducted in order to safely conclude these results. However, our proposed method is still a cheap one that can be applied in every level of the healthcare system.

4.4 Applications

The development of machine learning models using three distinct dataset - ultrasound only, biochemical only, and the combination of both - for Down Syn-

drome screening in both trimester 1 and 2 presents a versatile and practical approach to address real-world healthcare scenarios. This approach is particularly relevant for healthcare facilities that may face limitations in conducting comprehensive screening tests, such as ultrasound or triple tests. By offering a choice of three models, each tailored to a specific dataset, this strategy allows healthcare providers to adapt their screening methods to match their available resources and patient needs effectively.

While the utilization of a combined model incorporating both ultrasound and biochemical data in trimester 1 is ideal when there's a sufficient data from both tests, the option of relying solely on a model constructed from first-trimester biochemical data remains viable. This becomes especially relevant in scenarios where the healthcare facility lacks the capability to conduct Down's screening ultrasounds or when a patient opts not to undergo ultrasound testing. The accuracy of this standalone biochemical model closely aligns with that of the combined dataset model, offering a flexible choice of screening services.

In the second trimester, it's essential to perform both types of assessments due to the clear difference in accuracy between the model constructed from the combined dataset and the model built using two separate datasets. The models created from these two separate datasets could be used as a point of comparison for the current screening methods.

This new screening method is necessary in the current situation of Vietnam. Each year Vietnam has nearly 1.5 million new children born. Notably, each year there are about 1,400 -1,800 children with Down syndrome. In order to improve the quality of the race, Resolution No. 21-NQ/TU dated October 25, 2017 of the Sixth Conference of the 12th Central Committee of the Party on population in the new situation set the target for 2020: 70% of pregnant women and newborn mothers would be screened before birth and newborn, 90% of newborns would be screened for at least 5 common congenital diseases to overcome the situation of high number of children with birth defects. Up to now, the project on screening and diagnosing prenatal and neonatal

diseases developed by the Committee for Population - Family - Children has been implemented in 63/63 provinces and cities across the country. However, the implementation of this project also encountered many difficulties in terms of both medical personnel at health facilities as well as the awareness of the people about prenatal and postnatal screening. Only in 2017 alone, the Project carried out screening on 48.5% of pregnant women by ultrasound technique, and 29.7% of newborns were screened.

4.5 Study strengths and limitations

4.5.1 Study strengths

This study was the first study in North Vietnam to create machine learning based prediction models to screen for Down Syndrome. It also had a considerable number of pregnant women with Down Syndrome fetus, up to more than 1100 cases. This study worked on making screening methods for not only one but both trimesters. In addition, we also tried different combination of variables for different models.

4.5.2 Study limitations

The dataset used for training the AI model was collected by conveniently selecting samples from pregnant women undergoing prenatal screening at the Vietnam National Hospital Of Obstetrics and Gynecology. High-risk cases were either subjected to amniocentesis for fetal diagnosis or underwent genetic testing for the parents. While all cases were labeled to serve the software training, it's important to note that the software's recognition and differentiation capabilities couldn't be optimized entirely, especially for challenging cases. Although the Vietnam National Hospital Of Obstetrics and Gynecology is a major maternity hospital in the North with ethnic and regional diversity, the majority of cases were of the Kinh ethnic group and may not fully represent the entire Vietnamese population. Therefore, the initial machine learning models only "learned" from patients at the Vietnam National Hospital Of Ob-

stetrics and Gynecology and haven't learned from other groups of subjects. Additionally, the tests were conducted using equipment and techniques specific to the Vietnam National Hospital Of Obstetrics and Gynecology. In contrast, the test dataset was collected from two other hospitals, Hanoi Obstetrics and Gynecology Hospital and Hanoi Medical University Hospital. Differences in ethnicity, region of study subjects, testing techniques, machinery and equipment, reference thresholds, and other factors between these hospitals contributed to disparities in results between the training and test datasets.

5 Chapter 4: Conclusions

Based on the result above, we draw out 2 conclusions:

- Machine learning models built on combined ultrasound and biochemical data gave better results than a model built from two separate data sets in both trimester 1 and 2.
- The best model in trimester 1 was XTreme Gradient Boosting with 93% accuracy in validating phase and 100% accuracy in the testing phase. The best model in trimester 2 was Support Vector Machine with 89% accuracy in the validating phase and 99.9% accuracy in the testing phase.

6 Chapter 5: Recommendations

We want to give out some recommendations for these models:

- Since all these models were developed in R, they can be deployed on Shiny web app so that everyone can access it as long as they have a smart device.
- Then, these models could be tested with larger sample sizes, at more different hospitals to increase the diversity of Down syndrome cases, thereby generalizing the model in the population, increasing the model's screening ability in the entire population.
- After testing, test data could be applied as new training data and hyper-parameters can be re-adjusted, ensuring that models are regularly updated with new data.

For further studies:

- Further studies should include both ultrasound and biochemical data in the development of Down screening models.
- Data of future studies should be gathered from different sources, with the variables to be collected clearly designed and the entire clinical data collection process controlled to ensure data quality.

References

1. CDC. Facts about Down Syndrome | CDC. *Centers for Disease Control and Prevention*. Published online April 2021. Accessed March 28, 2022. https://www.cdc.gov/ncbddd/birthdefects/downsyndrome.html
2. Presson AP, Partyka G, Jensen KM, et al. Current estimate of Down Syndrome population prevalence in the United States. *The Journal of Pediatrics*. 2013;163(4):1163-1168. doi:10.1016/j.jpeds.2013.06.013
3. Park SC, Mathews RA, Zuberbuhler JR, Rowe RD, Neches WH, Lenox CC. Down Syndrome With Congenital Heart Malformation. *American Journal of Diseases of Children*. 1977;131(1):29-33. doi:10.1001/archpedi.1977.02120140031003
4. Roizen NJ, Wolters C, Nicol T, Blondis TA. Hearing loss in children with Down syndrome. *The Journal of Pediatrics*. 1993;123(1):S9-S12. doi:10.1016/S0022-3476(05)81588-4
5. Shott SR, Joseph A, Heithaus D. Hearing loss in children with Down syndrome. *International Journal of Pediatric Otorhinolaryngology*. 2001;61(3):199-205. doi:10.1016/S0165-5876(01)00572-9
6. Ram G, Chinen J. Infections and immunodeficiency in Down syndrome. *Clinical and Experimental Immunology*. 2011;164(1):9-16. doi:10.1111/j.1365-2249.2011.04335.x
7. Reilly C. Autism spectrum disorders in Down syndrome: A review. *Research in Autism Spectrum Disorders*. 2009;3(4):829-839. doi:10.1016/j.rasd.2009.01.012
8. McGrath RJ, Stransky ML, Cooley WC, Moeschler JB. National Profile of Children with Down Syndrome: Disease Burden, Access to Care, and Family Impact. *The Journal of Pediatrics*. 2011;159(4):535-540.e2. doi:10.1016/j.jpeds.2011.04.019

9. Boulet SL, Molinari NA, Grosse SD, Honein MA, Correa-Villaseñor A. Health care expenditures for infants and young children with Down syndrome in a privately insured population. *The Journal of Pediatrics*. 2008;153(2):241-246. doi:10.1016/j.jpeds.2008.02.046

10. Choi H, Van Riper M, Thoyre S. Decision Making Following a Prenatal Diagnosis of Down Syndrome: An Integrative Review. *Journal of Midwifery & Women's Health*. 2012;57(2):156-164. doi:10.1111/j.1542-2011.2011.00109.x

11. Erickson JD. Down syndrome, paternal age, maternal age and birth order. *Annals of Human Genetics*. 1978;41(3):289-298. doi:10.1111/j.1469-1809.1978.tb01896.x

12. Pandya PP, Brizot ML, Kuhn P, Snijders RJ, Nicolaides KH. First-trimester fetal nuchal translucency thickness and risk for trisomies. *Obstetrics and gynecology*. 1994;84(3):420-423.

13. Mersy E, Die-Smulders CEM de, Coumans ABC, et al. Advantages and Disadvantages of Different Implementation Strategies of Non-Invasive Prenatal Testing in Down Syndrome Screening Programmes. *Public Health Genomics*. 2015;18(5):260-271. doi:10.1159/000435780

14. Schiøtt KM, Christiansen M, Petersen OB, Sørensen TL, Uldbjerg N. The "Consecutive Combined Test"—using Double test from week 8 + 0 and Nuchal Translucency Scan, for first trimester screening for Down Syndrome. *Prenatal Diagnosis*. 2006;26(12):1105-1109. doi:10.1002/pd.1487

15. Shailaja K, Seetharamulu B, Jabbar MA. Machine Learning in Healthcare: A Review. In: *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*.; 2018:910-914. doi:10.1109/ICECA.2018.8474918

16. Neocleous AC, Nicolaides KH, Schizas CN. First Trimester Noninvasive Prenatal Diagnosis: A Computational Intelligence Approach. *IEEE journal of biomedical and health informatics*. 2016;20(5):1427-1438. doi:10.1109/JBHI.2015.2462744

17. Spencer K. Accuracy of Down syndrome risks produced in a first-trimester screening programme incorporating fetal nuchal translucency thickness and maternal serum biochemistry. *Prenatal Diagnosis*. 2002;22(3):244-246. doi:10.1002/pd.312

18. O'Callaghan SP, Giles WB, Raymond SP, McDougall V, Morris K, Boyd J. First trimester ultrasound with nuchal translucency measurement for Down syndrome risk estimation using software developed by the Fetal Medicine Foundation, United Kingdom–the first 2000 examinations in Newcastle, New South Wales, Australia. *The Australian & New Zealand Journal of Obstetrics & Gynaecology*. 2000;40(3):292-295. doi:10.1111/j.1479-828x.2000.tb03337.x

19. He F, Lin B, Mou K, Jin L, Liu J. A machine learning model for the prediction of down syndrome in second trimester antenatal screening. *Clinica Chimica Acta*. 2021;521:206-211. doi:10.1016/j.cca.2021.07.015

20. Neocleous AC, Nicolaides KH, Schizas CN. Intelligent Noninvasive Diagnosis of Aneuploidy: Raw Values and Highly Imbalanced Dataset. *IEEE journal of biomedical and health informatics*. 2017;21(5):1271-1279. doi:10.1109/JBHI.2016.2608859

21. Papaioannou M, Neocleous C, Schizas CN. Non-invasive Trisomy 21 Diagnosis Using Fuzzy Cognitive Maps. In: Kyriacou E, Christofides S, Pattichis CS, eds. *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*. IFMBE Proceedings. Springer International Publishing; 2016:731-736. doi:10.1007/978-3-319-32703-7_140

22. Khattak MT, Supriyanto E, Aman MN, Al-Ashwal RH. Predicting Down syndrome and neural tube defects using basic risk factors. *Medical & Biological Engineering & Computing*. 2019;57(7):1417-1424. doi:10.1007/s11517-019-01969-0

23. Neocleous AC, Syngelaki A, Nicolaides KH, Schizas CN. Two-stage approach for risk estimation of fetal trisomy 21 and other aneuploidies using computational intelligence systems. *Ultrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*. 2018;51(4):503-508. doi:10.1002/uog.17558

24. Catic A, Gurbeta L, Kurtovic-Kozaric A, Mehmedbasic S, Badnjevic A. Application of Neural Networks for classification of Patau, Edwards, Down, Turner and Klinefelter Syndrome based on first trimester maternal serum screening data, ultrasonographic findings and patient demographics. *BMC medical genomics*. 2018;11(1):19. doi:10.1186/s12920-018-0333-2

25. Koivu A, Korpimäki T, Kivelä P, Pahikkala T, Sairanen M. Evaluation of machine learning algorithms for improved risk assessment for Down's syndrome. *Computers in Biology and Medicine*. 2018;98:1-7. doi:10.1016/j.compbiomed.2018.05.004

26. Li L, Liu W, Zhang H, Jiang Y, Hu X, Liu R. Down Syndrome Prediction Using a Cascaded Machine Learning Framework Designed for Imbalanced and Feature-correlated Data. *IEEE Access*. 2019;7:97582-97593. doi:10.1109/ACCESS.2019.2929681

27. Islam SM, Pinki F. Partial View-Based Object Recognition using Leave-one- out Approach with Classification and Regression Trees. Published online January 2018.

28. Lynn B. Jorde JCC MJB. *Medical Genetics*.; 2019.

29. Locatelli A, Piccoli MG, Vergani P, et al. Critical appraisal of the use of nuchal fold thickness measurements for the prediction of Down syndrome. *American Journal of Obstetrics and Gynecology*. 2000;182(1):192-197. doi:10.1016/S0002-9378(00)70512-6

30. Carlson LM, Vora NL. Prenatal Diagnosis. *Obstetrics and gynecology clinics of North America*. 2017;44(2):245-256. doi:10.1016/j.ogc.2017.02.004

31. He F, Lin B, Mou K, Jin L, Liu J. A machine learning model for the prediction of down syndrome in second trimester antenatal screening. *Clinica Chimica Acta*. 2021;521:206-211. doi:10.1016/j.cca.2021.07.015

32. Benacerraf BR, Gelman R, Frigoletto FD. Sonographic Identification of Second-Trimester Fetuses with Down's Syndrome. *New England Journal of Medicine*. 1987;317(22):1371-1376. doi:10.1056/NEJM198711263172203

33. Nyberg DA, Resta RG, Luthy DA, Hickok DE, Williams MA. Humerus and femur length shortening in the detection of Down's syndrome. *American Journal of Obstetrics and Gynecology*. 1993;168(2):534-538. doi:10.1016/0002-9378(93)90487-4

34. Benacerraf BR. The Role of the Second Trimester Genetic Sonogram in Screening for Fetal Down Syndrome. *Seminars in Perinatology*. 2005;29(6):386-394. doi:10.1053/j.semperi.2005.12.003

35. Lockwood C, Benacerraf B, Krinsky A, et al. A sonographic screening method for Down syndrome. *American Journal of Obstetrics and Gynecology*. 1987;157(4, Part 1):803-808. doi:10.1016/S0002-9378(87)80059-5

36. Spencer K. Second trimester prenatal screening for Down's syndrome using alpha-fetoprotein and free beta hCG: A seven year review. *BJOG: An International Journal of Obstetrics & Gynaecology*. 1999;106(12):1287-1293. doi:10.1111/j.1471-0528.1999.tb08183.x

37. Souter VL, Nyberg DA, El-Bastawissi A, Zebelman A, Luthhardt F, Luthy DA. Correlation of ultrasound findings and biochemical markers in the second trimester of pregnancy in fetuses with trisomy 21. *Prenatal Diagnosis*. 2002;22(3):175-182. doi:10.1002/pd.278

38. Chen T, He T. Xgboost: eXtreme Gradient Boosting.

39. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

40. Nielsen D. Tree Boosting With XGBoost.

41. Alice Zheng AC. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. 1st ed. O'Reilly Media; 2018.

42. Tarawneh AS, Hassanat AB, Altarawneh GA, Almuhaimeed A. Stop Oversampling for Class Imbalance Learning: A Review. *IEEE Access*. 2022;10:47643-47660. doi:10.1109/ACCESS.2022.3169512

43. Cederholm M, Haglund B, Axelsson O. Maternal complications following amniocentesis and chorionic villus sampling for prenatal karyotyping. *BJOG: An International Journal of Obstetrics and Gynaecology*. 2003;110(4):392-399. doi:10.1016/S1470-0328(03)02091-3

44. Creasman WT, Lawrence RA, Thiede HA. Fetal Complications of Amniocentesis. *JAMA*. 1968;204(11):949-952. doi:10.1001/jama.1968.03140240005002