**Cover Page: Your contact info, project title, GPA, major(s), faculty mentor, estimated graduation date.**Contact Information: [ahoss1@unh.newhaven.edu](mailto:ahoss1@unh.newhaven.edu)  
Project Title: Research and Development of a Machine Learning Algorithm to detect Coronary Heart Disease  
GPA: 4.00  
Major: Computer Science  
Minor: Mathematics  
Faculty Mentor: Dr. Stephanie Gillespie  
Estimated Graduation Date: May 31, 2023

**Student Qualifications** **(5%) [maximum of 300 words]  
WHY SHOULD YOU BE SELECTED?  
Prepare a statement about your background, interests, prior research and skills (e.g., research methods courses, lab training) and other experiences that will contribute to the successful completion of the project.  Include what you expect to learn, such as skills you will need to acquire to conduct your SURF research project.  
Required: Upload a 1-2 page resume/bio/CV in Word (.doc or .docx) or PDF format.**

I would be an asset as a Summer Undergraduate Research Fellow because of the experience I have gained in communication, time management, and teamwork through leading and developing different projects over the past year. This included programming a robot in LabVIEW, creating a room layout in AutoCAD, and designing a model puzzle in Inventor. Additionally, in the winter break of 2019, I independently engaged in learning C programming through the coding platform HackerRank. Through trial and error, I learnt these applications and I explored the different features that would allow me to improve my designs and code. I would work persistently to correct the inaccuracies that I would observe and to increase the efficiency of the systems. From searching different resources, I would succeed in learning a particular aspect of the language/software which I would implement and then study the results.

In recent months, I worked diligently with Dr. Gillespie on the proposal. We would meet weekly to discuss the research materials that I would uncover, and the improvements required in the drafts. This involved searching through multiple journals to find the most appropriate material beginning with medical research in heart disease to the technical aspect in Machine Learning. With more in-depth research, I was able to develop a research methodology that Dr. Gillespie approved and considered to be feasible. Lastly, by reading the journals, I was able to improve my technical writing abilities over the course of one month.

To conclude, the project would require that I learn Python 3.8.1 and Weka 3.8.4 in respect to software to build the model. Regarding the data, I would learn about the different steps associated with building the model: data uploading, data cleaning, testing and evaluating. Overall, I believe by continuing my research I would be successful in acquiring the necessary skills.

**Current State of the Field** **(20%) [maximum of 500 words]  
WHAT DO SCHOLARS ALREADY KNOW OR BELIEVE ABOUT THIS TOPIC?  
In this section, make a strong case for the significance of your study. Describe the current knowledge and research findings. Explain how your study will build upon and/or extend the existing knowledge base. This is an important section that should provide sufficient background for reviewers who may or may not be familiar with your topic. Cite references (in-text), as appropriate, and attach the reference list.  
Optional: Upload references cited in Word (.doc or .docx) or PDF format.**

Heart disease, also referred to as Coronary Heart Disease (CHD), is defined as the condition when the heart does not receive an adequate supply of oxygen and nutrients due to the blood flow in the coronary arteries being restricted or blocked. This is most often due to the buildup of plaque in them. CHD results in angina (chest pain) and with further narrowing of the arteries may cause a heart attack [1].

CHD has been the leading cause of death for the past 15 years. It was responsible for 15.2 million deaths out of 56.9 million worldwide in 2016, and has been the leading cause of death in nearly all economy-income countries [2] because of the high rate of misdiagnosis and medical expenses for patients [3]. In 2017, the percentage of total deaths due to heart disease was 23.5% in the US alone [4]. These severe numbers support the urgency in developing cost effective means of detection and prevention of heart disease.

Currently with advances in technology, new methods for detection of CHD are being implemented. Machine Learning (ML) can be used to detect the disease in people by training computers to recognize patterns in patient data and medical profiles from patients affected and not affected by heart disease. The attributes currently studied, such as age, sex, number of vessels colored, resting ECG and many more, do not require extensive and expensive tests [5], and could lead to the patient receiving a dependable diagnosis quickly. ML could also provide the possibility of early detection of heart disease with reasonable precision in patients depending solely on the physical symptoms.

The majority of ML algorithms have been tested in their ability to predict CHD upon the Heart Disease Data Set, hosted by the UCI Center for Machine Learning and Intelligent Systems [6]. The dataset consists of 75 medical attributes in each of the four databases it contains; however only 14 attributes in one of the databases (Cleveland) have been used by prior researchers. This is due to there being fewer missing attributes and more records in the Cleveland database compared to other databases in the dataset [7]. Conclusions by Palaniappan and Awang suggest that incorporating those additional attributes and databases in the prediction systems could increase accuracy in predicting CHD [8].

The ML algorithms primarily used by researchers were Naïve Bayes, Artificial Neural Network, and Decision Trees [8-12]. However, multiple other types of ML algorithms have been developed - Support Vector Machine, K-Nearest Neighbor - that could be promising based on their success detecting other diseases such as diabetes and chronic kidney failure [12-14]. There is a current lack of exploration of these ML algorithms when using a larger number of attributes, and the effect it may have on overall accuracy. By using the larger data-sets with state of the art ML algorithms, we may be able to better learn patterns to predict CHD, such as characteristics of patients with CHD and the impact and relationship of medical attributes and CHD [8].

**Research Question/Hypotheses** **(20%)  [maximum of 500 words]  
WHAT IS THE PURPOSE OF YOUR STUDY? WHAT ARE THE RESEARCH QUESTIONS? What is the overarching goal or purpose of your study? What research question(s) do you propose to investigate?  What is your hypothesis (if applicable) and what do you expect to find?  What are the expected outcomes (e.g., What will your project contribute to the field or discipline?)**  
The purpose of the study is to develop five essential machine learning classification techniques/algorithms that would predict heart disease with reliable accuracy and to identify the optimal number of attributes required to produce a comprehensive diagnosis, using the Heart Disease Data Set of the UCI repository (consisting of four databases: Switzerland, Switzerland, Cleveland, and Budapest), for each classification technique. Since, there is a lack of knowledge about the effect of the attributes and databases not considered in previous research, this would provide information as to the effect on the accuracy of the model and how it would change with the addition of other attributes. The research questions being investigated here are primarily the accuracy of the different models that would be developed for predicting heart disease and the optimal number of attributes required for each model. Finally, using the complete dataset and attributes, patterns in the data that would emerge from the model would be used to answer the following questions:

1. The order of the most significant medical attributes being inputted in the model
2. Impact and relationship of the medical attributes in predicting heart disease
3. Characteristics of patients with heart disease
4. The accuracy of the models in predicting heart disease in patients for a given medical profile

After developing the models, three classifiers with the highest accuracy would undergo hyperparameter tuning so that they may be optimized further allowing more patterns and relationships to emerge from the data, hence, resulting in three highly optimized models for predicting CHD in patients.

Expected Outcomes:  
It may be observed that due to the large number of attributes, the accuracy of the predictive model may be low, however, the effect and relationship of attributes and databases not considered in previous research would allow new patterns to emerge. Also, certain patterns and relationships may have to be ignored after careful medical evaluation as those patterns may be random. Lastly, the accuracy of the primary models (Naïve Bayes, Neural Network, and Decision Tree) are expected to be within 75-95% [8,9,10,11,12] and the secondary models (Support Vector Machine and K-Nearest Neighbor) to be within 70-85% for this experiment [12,13,14]. We also hypothesize that fewer attributes would result in more accuracy however it would result in a loss of knowledge about the data not being considered.

**Research Design and Methodologies (25%) [maximum of 600 words]  
WHAT IS THE RESEARCH DESIGN AND WHAT METHODOLOGIES WILL BE USED?  
Describe the research design and why that design is appropriate for the research questions and/or topic you are investigating. Describe how you will collect and analyze data. If applicable to your project, describe how you will test your hypothesis or prototype. You may attach instruments (e.g., survey or interview questions) or other supporting documentation (e.g., images, technical specifications, design criteria) that will demonstrate to a reviewer that the research design and methodologies are appropriate for your proposed study and adequate for accomplishing your project goals.  
Optional: Upload a supplemental document (e.g. survey questions, technical specs, images, etc.). Acceptable files include: Word (.doc or .docx), Excel, PDF format, GIF, and PNG.**

**Establishing a Baseline**

The Heart Disease Data Set is open source and freely available at the online UCI repository. Initially, the different databases in the Heart Disease Data Set would be cleaned to remove missing values. For this experiment, the primary focus would be the Cleveland database consisting of 75 attributes and 303 instances, however the remaining databases would also be explored. The cleaned Cleveland dataset would be divided into 3 subsections approximately 70% training dataset, 20% testing dataset and 10% validation dataset. The training dataset would be uploaded on Weka 3.8.4, an open source Machine Learning (ML) software which consists of a graphical user interface that allows users to visualize datasets and compare classifiers. The primary classifiers (Naïve Bayes, Neural Network, and Decision Tree) would be applied to the dataset followed by the secondary classifiers (Support Vector Machine and K-Nearest Neighbor). The resulting models in Weka would then be applied on the testing dataset to evaluate the performance of the classifiers, in terms of accuracy, specificity, sensitivity, and F – measure. The results would form a baseline value for the classifiers.

**Feature Selection**

After establishing the baseline, the optimal number of attributes would be selected for each classifier. This is done using an iterative process and the Attribute Selection Feature in Weka. The feature ranks the attributes respect to their importance in predicting the class. Additionally, using the attribute subset evaluation feature, groups of attributes that would be most reliable in predicting the class would be determined. After doing so the dataset would be modified to remove the attributes that have a lower predictive power and would be divided again into training and testing datasets. The classification and evaluation process in the previous step would be repeated. The corresponding performance values would be recorded and a graph of each classifier (Performance against Attributes) would be plotted. After multiple iterations it is expected that the accuracy would initially increase and then decrease like a bell curve for each classifier.

**Conclusions with Weka 3.8.4**

When using Weka, k-fold cross validation will be used on the training datasets to obtain an average value for the performance evaluators. Based off the top features selected, the three classifiers with the highest accuracy and minimum (optimal) number of attributes would be selected. Additionally, in Weka, using the Attribute selection feature, the order of the most significant medical attributes can be determined. Using ML algorithms, the impact of heart disease and characteristics of patients with heart disease can be reasonably concluded.

**Supplemental Conclusions with Pythons 3.8.1**

By this time we would have the dataset with the optimal number of attributes for each classifier. If time permits, we will build the three classifiers selected in the previous section in python ML and use hyperparameter tuning to optimize them further. For example in k nearest neighbor, the value of k can be varied to find the model with the highest accuracy by using a “for loop”. For different classifiers, the hyperparameter tuning would be different. Also depending on the classifier and the data we have gathered; we would implement other tuning techniques that would be based more on intuition and experimentation. After doing so, new patterns that would emerge would be recorded. The two non-hyper tuned classifiers in Weka and the three hyper tuned classifiers in Python would be tested on the validation dataset. The resulting model with the highest accuracy and the least number of attributes would be considered as the most reliable model for predicting CHD in patients.

**Timeline** **(5%) [Please provide a bulleted list or table.]  
WHAT IS THE PROPOSED TIMELINE FOR STARTING AND COMPLETING THE PROJECT?  
Please provide a timeline of tasks and activities reflecting your plans for completing the project by the end of the SURF period (July 31, 2020).  Include any pre-program required tasks such as IRB/IACUC applications, CITI training, lab safety training, and/or ordering initial supplies.  This section should demonstrate that you are prepared to complete the project within the ten weeks of SURF.  It will also assist you and your mentor in planning. Be sure to include planning meetings with your mentor, especially during the first three days of SURF 2020 (May 27 -29) as well as mandatory seminars (June 1 - 3) and final presentations and seminars during the last week of July.**

|  |  |
| --- | --- |
| **Tasks/Activities** | **Dates** |
| CITI Responsible Conduct in Research Training Due | May 01, 2020 |
| Orientation Session “How Things Work” | May 01, 2020 |
| Planned Meeting with Mentor \* | May 18, 2020 – May 22, 2020 |
| Weka 3.8.4 and Python 3.8.1 download. Heart Disease Data Set download.  Begin dataset cleaning. | Week 01 (May 27, 2020 – June 01, 2020) |
| Surf Seminar | Week 01  (June 01, 2020 – June 03, 2020) |
| Finish dataset cleaning.  Upload dataset in Weka. Begin establishing baseline. Planned Meeting with Mentor \* | Week 02 (June 03, 2020 – June 10, 2020) |
| Complete establishing baseline.  Begin working on feature selection. Planned Meeting with Mentor \* | Week 03 (June 10, 2020 – June 17, 2020) |
| Continue feature selection.  Begin answering hypothesis questions.  Planned Meeting with Mentor \* | Week 04  (June 17, 2020 – June 24, 2020) |
| Complete feature selections.  Answer hypothesis questions.  Select top three classifiers. Planned Meeting with Mentor \* | Week 05  (June 24, 2020 – July 01, 2020) |
| Check in with the SURF staff | Week 05  June 2020 (TBD) |
| Build classifiers in Python.  Begin hyper tuning classifiers Planned Meeting with Mentor \* | Week 06  (July 01, 2020 – July 08, 2020) |
| Complete hyper tuning classifiers. Planned Meeting with Mentor \* | Week 07  (July 08, 2020 – July 15, 2020) |
| Test on validation set and finalize results Begin working on Poster and Final Report Planned Meeting with Mentor \* | Week 08  (July 15, 2020 – July 22, 2020) |
| Finish SURF Poster and Final Report Preparing for Oral Presentation Planned Meeting with Mentor \* | Week 09  (July 22, 2020 – July 29, 2020) |
| Surf Seminar | Week 10  (July 29, 2020 – July 31, 2020) |
| Surf Research Symposium and Poster Day | September 2020 (TBD) |
| Final Report Due | September 2020 (TBD) |

\*Planned Meeting each week Monday morning and Wednesday afternoon.

**Impact on Student Success** **(5%)  [maximum 200 words]  
HOW WILL PARTICIPATION IN THE SURF PROGRAM CONTRIBUTE TO YOUR FUTURE SUCCESS?  
Describe how your participation in the SURF program will contribute to your personal and/or academic success. That is, how do you expect to use what you learned in future coursework, graduate school, and/or careers?**

I intend to work in a technical job focused on Artificial Intelligence, specifically Machine Learning (ML), and this research project would provide me with invaluable insight into the field. I would be able to gain knowledge about the different technical aspects related to ML, and there would be an increased chance of me obtaining a respectable position in a company as they would appreciate the prior hands-on experience that I have gained. Additionally, over the course of these two months, I was able to develop a passion for researching and learning new materials through reading different journals. I could understand the value of the systematic processes that were being employed for each research paper, through rigorous discipline, to ensure an unbiased and dependable result. I was able to appreciate the value of the knowledge being shared among researchers to bring about a cumulative growth in a field to answer complex questions. This program would let me explore these different aspects of research and conceivably consider a future in doctorate studies. Lastly, by learning new programming languages and applications, I would be able to continue to develop different projects over the course of 3 years and explore the field further.