Computer Science Capstone
CS 476 - Requirements Engineering

SmartState

Project Title: Event-driven MachinE Learning, Intelligent Assessor (EMELIA)

Technology Feasibility Analysis

Overview:

The purpose of this document is to present this project at a more technical level, in addition to showcasing our technology and integration challenges. This document will also go more in-depth to the evaluations that underwent as a team and the solutions that were made.

Team Members:
Andrew Hurst
David Rodriguez
Reed Hayashikawa
Jianxuan Yao
Jesse Rodriguez

Clients:
Rick Duarte
Jon Lewis

Capstone Mentor:
Fabio Marcos De Abreu Santos

Northern Arizona University
School of Informatics, Computing, and Cyber Systems

General Dynamics
Mission Systems

Table of Contents

1. Introduction

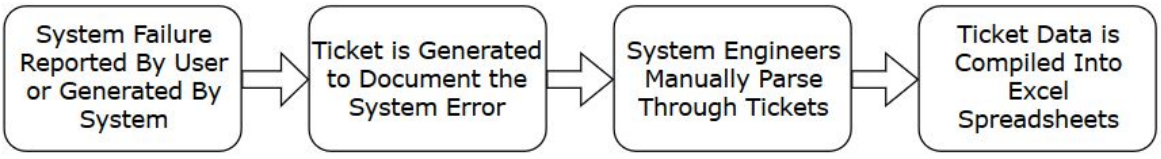
General Dynamics Missions Systems is a global aerospace and defense company that develops C4ISR solutions in all areas such as land, air, and cyber domains. C4ISR stands for Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance systems. General Dynamics develops and maintains and advanced command, control and direction-finding communications system to execute search and rescue missions. These search and rescue missions are conducted by the U.S. Coast Guard with the use of this advanced communications system. The name for the partnership project between General Dynamics and the U.S. Coast Guard is Rescue 21.

The current communications system used by Rescue 21 relies on a ticket system to generate reports regarding system or hardware failures. These tickets contain data regarding:

- Code identifier for the error
- Length of time for which the error was present
- Location for where the error occurred
- How the error was received
- Description of the error
- Description of what was impacted by the system error
- Information regarding the resolution for reported error
- The component impacted by the system failure

Each error must be classified according to the error type to assist system engineers in maintaining the Rescue 21 communications system. System engineers manually parse through the tickets in order to classify the type of error reported. This process is inefficient and reduces our client's ability to respond to system failures.

Figure 1



Our sponsor, General Dynamic Missions Systems has asked our team to create a solution to improve the Rescue 21 communications system. Rick Duarte and Jon Lewis are two systems engineers that will oversee our project. Rick and Jon are part of a larger team of engineers that oversee all of the complex communications systems for our client. Each of these systems are highly specialized and require time and resources for maintenance and enhancement. By improving system operations and reducing outages, our client is able to better serve the U.S. Coast Guard and other systems that support the U.S. Department of Defense.

Our solution is to develop an Event-driven MachinE Learning, Intelligent Assessor (EMELIA) to classify the system failures for our client. EMELIA will be able to :

- Take in data created from the ticket system as input
- Pass the input to a neural network to begin the classification process
- Classify the input data with a 90% success rate
- Utilize a command-line interface for use by engineers
- Accept commands that will assess EMELIA's data classification performance

This document is meant to explore the technologies that are available to implement our solution. A feasibility analysis will reveal the most practical technologies that will allow us to make decisions regarding the design, implementation, testing, release, and maintenance for our machine learning classifier, EMELIA.

The document will be organized into three main categories. The first section will highlight the challenges that our team anticipates during the implementation of EMELIA. This section focuses on challenges regarding framework selection, storage of classified data, version control for source code, issue tracking, and task management, and testing harnesses. The following section is the analysis that will explore alternatives that are presented in the previous challenges section. This section will be a collection of information that details the technological tools and how each tool contributes to EMELIA's implementation. The last section will detail the integration of the tools evaluated in the analysis section, and the final determination for the tools that will be used to build our machine learning classifier. The technological choices will be metric based and be decided based on the four main categories: Cost, Usability, Features, and Documentation.

2. Technological Challenges

Below are all of the various technologies and challenges we need to analyze in order to execute a solution that will satisfy and meet the requirements for our client. Each challenge were selected based on their contribution to the technological feasibility for our envisioned solution. The challenges include the classifying model, machine learning framework, version control, project management, and programming language that will be used to build the project.

2.1 Classifying Model

In order to properly execute our machine learning solution for our client, we need to choose a classifying model that will best integrate with our project's technologies. The model we choose will have to be able to process data efficiently and in a timely matter. The model we will choose will need to meet these requirements and contain features that will be directly useful for our project. We will judge each model based on criteria including:

- Accuracy
- Variability
- Performance

. The main challenge for this technology is choosing a model that can assist with the proper machine learning algorithms for our project and have features that will directly work towards the project's goal. As stated in the introduction, we are taking a series of input formatted in a CSV file. This file will then contain several columns and rows of data representing individuals tickets which we will then extract the data in order to train our model and eventually achieve a 90% accuracy in classifying and identifying ticket types. There are many models in neural networks that process images and or use regression in order to statistically make assumptions and choices. With that being said, we need a machine learning model that can be able to process input data in the form of text and be able to classify and save this in memory for later analysis. This will require our team to explore the different machine learning classification models and identify features that will increase our success.

2.2 Framework

We need a dynamic framework as the foundation for our application and other technologies. The framework we will choose for our project will have to contain attributes that can assure the effectiveness and efficiency of this project.
Here are some attributes we are looking for in a framework:

- Language
    - Familiarity with all members is preferred.
    - Small learning curve.
- Scalability
    - General Dynamics is looking for a product that is maintainable and will evolve with the organization.
    - Easily scalable to adjust as the project progresses.
- Capability
    - Easy to integrate, execute, and maintain.
    - Able to send data in a uniform format to other sectors in the client's organization.

The right framework will assist us by providing a wider scope of the program's functionality and give us the tools to selectively manipulate the framework to address our challenges. We decided to list frameworks as one of our technological challenges because the complexity of encapsulating our language and providing useful tools is crucial to the success of this project. With that being said, in order to fulfill our client's request we need to be prepared for challenges, thus we will need a framework that is easily understandable and maintainable if refactoring is needed. Without the proper framework, implementing this project will be more difficult and will present delays that could be easily avoided. Thus, we will need to choose a framework that best fits our needs and directly assist us in implementing this project.

2.3 Version Control

As a team, we need a way to collaborate on the project without conflicting issues.Version control is a system that records changes in the contents of one or more documents for future reference of specific version revisions. With it, you can trace a file back to its previous state, or even the whole project back to its previous state at a certain point in time. You can compare the details of changes in the file, find out who modified where in the end, find out the cause of the weird problem, and who reported a functional defect, etc Using a version control system usually also

means that even if you make a mess of changing, deleting and deleting files in the whole project, you can still easily restore to the original But the extra work is minimal. For version control, we need to choose a method that can incorporate the following:

- Accessibility
    - We want a simple to use application that we can pull, branch and push content.
- Management
    - We want an organized repository for code reviews and pull requests instead of manually merging content in order to work cohesively.
- Arrangement
    - We need to fill the whole process with a sense of hierarchy, which helps us to divide the stages of the project easily. It can make our compilation clearer.
- Necessity
    - Version control systems are systems that record changes in the contents of one or more files for future reference of specific version revisions.
    - This system can automatically help us to back up every change of files, and can easily restore to any backup (version) state.
- Importance
    - Without version control system, the code may be accidentally overwritten or lost by others or themselves, who changed the code for any reason, and there is no way to recover the changes of the previous days. With version control system, developers only need to record every code change and update it through version control system.
    - With the version control system, we can browse all the development history records, master the team's development progress, and make any changes are no longer afraid, because you can easily restore the previous normal version. We can also distribute different versions of the software through branch and tag functions, such as stable version, maintenance version and under development version.
- Detail
    - Frequent submission can help the team to make the development progress transparent and reduce the codes' conflicts when multiple people develop. When multiple people modify the same piece of code at the same time, it is very troublesome to solve code conflicts. In addition, we also hope that each submission has an appropriate granularity, that is, each submission should have a high degree of relevance and independence. For example, a small function or improvement. If you're doing both a and bug fixes, you should submit them separately. Grammatically incorrect programs that cannot be constructed should not be submitted, causing team problems.

    - The advantage of having a good code submission habit is that in addition to a clearer version history, we can easily restore or migrate the code. Assuming that there is a problem with the above a new function , we can restore only a without affecting the repaired bug, or select only the selected bug to migrate to different branches.

In order for our project to be implemented efficiently and optimally, we will need to utilize a version control software. This version control software will be directly linked to the development steps of our project and will have a tremendous impact on our project's progression. The ability of our team to work together effectively with the least amount of conflicts is very important. The most common attribute that exists in all version control software is the ability to retrieve files and be able to update the content of a file in real-time. The system records changes to a file , or set of files over the course of development, so that specific versions can be recalled if necessary. However, we will be looking for other attributes that can assist us in maintaining our code in the simplest way. Since every team member will be making changes to the files of our project, these changes will possibly conflict as the project progresses. Therefore, we will need to find a version control software that can best assist us in making this process less of an issue. The right choice in software should resolve version control issues and allow the optimal implementation of our project.

2.4 Progress Management

We will need to manage our project in the simplest way possible. With this project going into the next year, we will have several requirements to fulfill, components to modify, bugs, and many other obstacles. We need a way to monitor the progression of our project and below are some of the challenges that exist in choosing a service:

- Cost
    - The development team would prefer to utilize an open-source or free software to keep costs at a minimum.
- Familiarity
    - The software should be familiar with all team members to make it easy to integrate into our current workflow.
- Usability
    - We need a service that will be able to classify issues, track the progress, and be able to close the issue when completed.
- Features

    - Features include the ability to assign tasks to team members, track progress on task completion, and integrate tools used by the team.

As we progress through the stages of development of our project, we will need to utilize an efficient software to track our progress and classify potential issues. The ability to address these issues and possibly predict future challenges is crucial to our development and the quality of our product. The cost of the software should not be too demanding for the development team to avoid delays in task completion. If the software is familiar with the entire development team, we can efficiently integrate it into our current workflow. If the overhead for learning how to use the software is too high, then progress in the development of our project could be delayed. The project management software that contains the features stated above will assist in providing a more clear overview and a sense of direction for the project. Choosing the right software will assist the development team to prioritize issues and allow this project to move forward effectively.

2.5 Language

For a data classification machine learning project, choosing a language is not so obvious. There are several considerations when choosing a language for machine learning:

- Usability
    - The language comes with the lowest overhead in terms of a learning curve and setting up a development environment.
- Features
    - Features include available libraries, especially related to data science and computational capability.

- Framework Integration
    - The language we choose must integrate well with the machine learning frameworks the development team has analyzed for this project (Section 3.2).
- Documentation
    - A technologically feasible language must have comprehensive documentation to support various implementation methods.

Choosing a language is a technological challenge due to the implications of using the wrong language during the implementation stage. If the overhead for learning a language is too high, it will prevent the team from making the necessary progress. If the language does not support the functionality need to make necessary computations, the development team may be forced to stop implementation and delay deadlines. If a language cannot properly integrate with an established

machine learning framework the project will become technologically unfeasible. If a selected language lacks documentation that may assist the development team, then it can cause delays or disrupt the work distribution amongst the team. Choosing a technologically feasible language is a critical step in the development step of this project.

3. Technological Analysis

3.1 Classifying Model

Some basic qualifications of the model can be assumed in order to narrow down the options. The most obvious feature being classification into multiple categories while using a neural-network (request of the client). It is also known as the data to be classified are text-based documents. The client possesses a large data set of classified tickets that can be used to train the desired model. This narrows our research down to supervised document classification neural-network models. Each model is compared based on their cost, usability towards our project, features that directly benefit our group, and the supporting documentation it contains for further assistance.

3.1.1 Multilayer Perceptrons (MLP)

This neural network model uses input parameters, hidden layers, and an output layer to make predictions for a learning model. This model is well suited for classification problems where inputs are assigned a class or label. MLP's utilize backpropagation to compute the output error. The computation determines how much each neuron in the previously hidden layer contributed to the overall value. The output error is then passed in reverse through hidden layers, to adjust input parameters. Once the input parameters have been adjusted, the new parameter is sent through the learning model for re-evaluation.

In addition to the useful tools this neural network provides, it also contains characteristics that directly benefits our project. This neural network is well adapted for tabular datasets and being that our data will be in a CSV file this attribute directly relates to our project. Furthermore, this network is mainly used for classification and prediction. MLP is also known to have a flexible training model that can be applied to a wide range of data types. Although flexibility can reduce the level of complexity for handling a specific data type, the utilization of perceptron neurons increases efficiency and accuracy regarding classification problems. More information on this model is explained below[16]:

- Cost - Learning models do not come at a cost to development teams. Most of the frameworks used for machine learning come with functionality that supports learning models.

- Usability - This particular model is extremely easy to use within a TensorFlow framework. This model is very easy to visualize for the development team when using the TensorFlow tutorial. There are several optimization modules such as Keras that utilize other machine learning modules to increase performance and reduce complexity.

- Features - As mentioned previously, this learning model is easy to use because of the number of integrated features that support this particular learning model. Some of the features provided for this learning model are framework dependent, so the final choice for the framework section will be critical.

- Documentation - Documentation for this learning model includes video tutorials and written documentation. There is a lot of online content that will support the development team as we implement this learning model. More importantly, the documentation is framed according to classification problems.

3.1.2 Recurrent Neural Network(RNN)

This neural network allows previous outputs to be used as inputs and does not restrict the size of the inputted data. In addition, this network is proven to be most successful when working with long sequences such as in natural language processing. This network is designed to use sequence prediction problems. The problems can be classified as One-to-One(1 input - 1 output), One-to-Many (1 input - many output), and Many-to-many. This is useful in the option that we want to either classify an individual ticket to one or many different categories.

Although there are many applications for this network, there are attributes that are not directly related to our project and may cause further complications. For example, RNN is mostly used in the fields of speech recognition and language processing and does not work well with tabular data sets such as a CSV file or spreadsheets. In addition, the output of RNN is more dependent on the final input and may cause us to run into accuracy problems when training. As time goes by, it reduces the effect of the earliest inputs which is not a supporting attribute for our project and RNN has difficulties remembering information that is too far apart, which is a disadvantage of simple deep learning.
Some advantages of this network are that the model size does not increase along with the size of the input and that historical information is taken into consideration when making computations. However, this network is known to have its downsides. For instance, computations are known to

be time-consuming, and the network can have difficulties accessing information being that there is no restriction on its data size. More information on this model is explained below[15]:

- Cost - This learning model does not come at a cost.

- Usability - The possibilities and advantages when using this model are tremendous however can lead to issues in the future being that this model is geared mostly for the fields of speech and language processing.

- Features - As mentioned previously, this learning model has many advantages even in terms of features, however, the same problem of different implementation purposes exist.

- Documentation - Documentation for this learning model includes video tutorials and written documentation. There is a lot of online content, however, relates back to the issue where most of these documents do not relate to what our team is trying to implement.

3.1.3 Convolution Neural Network(CNN)

The field in which these neural network covers are image recognition, image classification, object detection, facial recognition and much more. This system is very similar to other neural networks, however, this model uses a feed-forward type of architecture. This means it utilizes multilayer perceptrons which is designed to optimize preprocessing instead of utilizing its internal memory to analyze sequence of inputs. This model is known to be computationally efficient and effective for image processing and analysis.

We found that there exists some attributes that this network has that can benefit this project such as the ordered relationship between words in a document and the ordered relationship in time steps in which CNN uses the latter. In addition, convolutional neural networks is used to extract the feature of the target through a filter. First, the low-level features, eventually making its way to the high-level features. It then convolute them together to output. This can directly assist us by extracting the features of the target accurately in order to classify them.

Although this network contains many useful features, the main use for this network is used for image data and two-dimensional input. CNN also has a big disadvantage that is translation invariance. When two things have a certain feature, they can't distinguish the location of the feature. As long as they have the same feature, they will classify and treat the information the similarly. Another disadvantage is the pooling layer. The existence of pooling layer will lead to the loss of valuable information and will also ignore the relationship between the whole and the

part. Additional information that will be used to make a decision in ranking this model is explained below[14]:

- Cost - This learning model does not come at a cost.

- Usability - Similar to the other models, the possibilities and advantages when using this model is also a great benefit however this model is also geared towards a different implementation. Although this model is known to be computationally efficient, its main use is for image processing and analysis which is far from what this team is trying to implement.

- Features - As mentioned previously, this learning model has many advantages even in terms of features and attributes, however, the team's implementation goals cannot be reached with this modeling.

- Documentation - Documentation for this learning model includes video tutorials and written documentation. There is a lot of online content, however, relates back to the issue where most of these documents do not relate to what our team is trying to implement.

Table 2

	Cost	**Usability**	**Features**	**Documentation**	**Choice**
Multilayer Perceptrons	✔	✔	✔	✔	✔
Recurrent Neural Network	✔	✖	✔	✔	✖
Convolution Neural Network	✔	✖	✔	✔	✖

Conclusion

After extensive research in the many segments of neural network, our team has made an agreement to implement this project using the multilayer perceptron model. Given the dominant score, we are confident that this model will serve us well based on the criteria involving cost, usability, features, and documentation. Being that this model is suited well for classifying

prediction problems, adapted for tabular datasets, has flexible training models for various types of data, and has an increased advantage for accuracy and efficiency, we decided to give this model a perfect rating.

3.2 Framework

We will need to agree on a framework that has specific features and attributes. Attributes such as performance, scalability, and abstraction are crucial to our success. Below are some of the options that have been explored as potential machine learning frameworks.

3.2.1 TensorFlow

This is one of the most popular frameworks when producing artificial intelligent systems and networks. TensorFlow is an open-source machine learning platform that utilizes data flow graphs in order to carry out numerical computations. It consists of high-level APIs and flexible ecosystems of tools. It does not seem to have a steep learning curve and can be easily integrated into our system. TensorFlow offers a lot of resources to create and design high-level architectures for neural network systems. TensorFlow offers resources such as high-level APIs such as Keras which allow for faster prototyping and production. [4] Some things TensorFlow offers as a whole includes:

- Cost - Free open source software
- Usability - Has a simple, consistent interface that provides clear feedback for user errors
- Features - Functional APIs, core packages and many other features
- Documentation - Tensorflow offers a lot of documentation for training and evaluating ML models. It assists the user by providing guides for both developing the software and modeling the neural network.

3.2.2 SciPy

SciPy, which is a large ecosystem of open source software for scientific computing in Python. The SciPy stack contains a small number of core packages that include:

- NumPy as a fundamental package for numerical computation
- SciPy library which contains a collection of numerical algorithms
- Matplotlib as a plotting package that provides 2D and 3D plots

- Python as a programming language

The SciPy ecosystem includes general and specialized tools for data management and computation. Key packages for data computation include Pandas for high-performance easy-to-use data structures and data analysis tools and Scikit-learn as a collection of algorithms and tools for machine learning. Productivity and high-performance tools include IPython, Jupyter notebook, Cython, and IPyParallel. [4]

- Pandas organizes data into highly organized data structures. Pandas can receive various file formats and convert them into a DataFrame object in Python. Using the Pandas API available for DataFrame objects, a developer can make computational function calls on data stored within a Pandas DataFrame. This data can be filtered according to specified programming conditions, which would prove useful when comparing a test dataset to a training dataset within EMELIA. Another powerful piece of functionality using Pandas is its "Data Cleaning" capability. This capability allows for the removal of potential null values in a dataset that may offset expected output calculations. Since neural network models rely on weighted values for input data, avoiding null values will increase the speed and accuracy of our classifier. [7]

- Scikit-learn is of interest because of its specific application toward our project. The machine learning solution that we intend to build will be reliant on the classification functionality that scikit-learn provides. Within the classification module, scikit-learn contains a set of neural network models capable of classifying data utilizing supervised learning. Supervised learning is the process of learning a function that maps an input to an output based on example input-output pairs. Supervised learning is the training model that will be utilized to classify the system failures in ticket reports for our client. Scikit-learn contains classes focused on classification, regression, and regularization of data, which are part of the neural network machine learning model we intend to implement. [8]

- IPyParallel provides an architecture within IPython for parallel and distributed programming. The performance will be a concern regarding the performance of the classifier (EMELIA) that is being built for our client. Parallelization of EMELIA may be needed to meet the performance requirements specified by the client in the project proposal. IPyParallel provides algorithms to make programs fast and efficient. The classifier that our group intends to build will first optimize sequential programming techniques before there is an attempt to make our classifier run in parallel. [5]

- Cython is a static compiler that is part of the SciPy stack. Cython enables easy writing of C language extensions. This compiler is a superset of the Python language that supports calling C functions and declaring C types for variables and class attributes. Cython interacts well with large data sets using multi-dimensional NumPy arrays. The most essential part of this SciPy tool is its ability to integrate natively with existing code and data from low-level or high-performance libraries and applications. Cython's use in EMELIA would pertain to issues regarding integration with our client's source code. If the source code contains C programming language implementations within the current framework, we can neatly wrap our functionality with Cython to make calls to functions written in the C language. [1]

- IPython is an interactive shell for writing and executing Python code within the terminal or command prompt. IPython allows for code to be imported into a shell so that developers can make alterations and test source code in real-time. This tool supports interactive data visualization and use of GUI toolkits. IPython contains high-performance tools for parallel computing, which uses the IPyParallel package mentioned previously. This highly integrated system will allow for benchmarking of function calls throughout development if the client specifies parallelization as a requirement for EMELIA. [3]

- Jupyter notebooks is part of a larger tool known as JupyterLab. JupyterLab is a web-based interactive environment for Jupyter notebooks, code, and data. Jupyter notebooks support a range of workflows suited for data science, scientific computing, and machine learning. Jupyter notebooks allows developers to create and share documents that contain live code and visualizations. This tool can serve the development team by acting as a medium to run the classifier and produce helpful visualizations to ensure that the classifier is effectively determining the correct classification of data. The usefulness of This tool extends beyond the use of the SciPy environment. It can be used with various programming tools and with the TensorFlow machine learning library. [6]

SciPy is a sophisticated machine learning ecosystem. This development ecosystem has high integration with several data computation and high-performance tools that support the machine learning classifier we intend to engineer. This machine learning ecosystem will be rated on a 1 – 10 scale for cost, usability, features, and documentation.

- Cost - Since SciPy is an open-source, it will be free to use by the development team. It will reduce any overhead during development and will be available to all members of the team throughout development.

- Usability - After looking through the documentation for some of the computational tools, SciPy can be downloaded very quickly and used immediately in a Python project. This product can be easily installed with a Conda development environment. Utilizing a Conda environment to manage package dependencies is part of the rating for usability.

- Features - As outlined above, there is a complete set of tools that make this project feasible from a technological standpoint. This highly integrated system can be very useful for machine learning, as long as there is no human involvement during the process of classifying data. If deep learning is a requirement specified by the client to classify datasets, scikit-learn will be inadequate.

- Documentation - Documentation related to the tools supported by SciPy are sparse in terms of examples. Due to the inexperience within our development team and machine learning, it is necessary to have working examples of code implementation.

3.2.3 Pandas

Pandas is an open-source library that extends the functionality of python to include the manipulation of data sets that are able to be imported from various sources such as CSV, Microsoft Excel Spreadsheets, and SQL databases. While not specifically related to machine learning, Pandas enables us to extract and prepare data before training, which would be helpful for an accurate machine classifier. Pandas use DataFrames and Series to intelligently manage data sets and contain a variety of tools that allow the DataFrame to be flexible. Pandas is popular, easy to learn, and contains extensive documentation. Pandas is compatible with other packages such as scikit-learn, making it easy to integrate into existing systems. Pandas do not implement models outside of simple linear and panel regression, so more sophisticated modeling should be left to other tools such as scikit-learn. Pandas is optimized with code paths written in Cython or C. [13]

Pandas offers a variety of features to manipulate data sets including:

- Manipulation - moving columns, cutting, reshaping, merging,
- Time-Series Handling - operations on time/date, resampling, and automatically aligning data sets
- Missing Data Management - automatically exclude dummy values, drop null values, replace bad values, or interpolate missing values
- Grouping Operations - group by similar to methods used in SQL Databases
- Summary Statistics - offers a fast summary of data statistics
- Visualization - access to simple plots on data structures

This machine learning ecosystem will be rated on a 1 – 10 scale for cost, usability, features, and documentation.

- Cost - Pandas is open-source, so the cost to the development team is zero.
- Usability - Pandas can be integrated easily in a Python project. Pandas has limitations for data modeling, so other tools might be more helpful.
- Features - Pandas has a lot of features for data manipulation, but not much else.
- Documentation - Pandas has extensive documentation that is updated as new features come out.

Table 3

	Cost	**Usability**	**Features**	**Documentation**	**Choice**
TensorFlow	✔	✔	✔	✔	✔
SciPy	✔	✔	✔	✔	✖
Pandas	✔	✔	✖	✖	✖

Conclusion

Overall, the deficiency SciPy presents to the development team involves its inability to utilize scikit-learn to classify datasets using deep learning models. This tool best serves projects that focus on a supervised learning training model. It will be more useful if the system only requires a comparison of test data to predefined, and organized training data. If machine learning is the goal, SciPy utilization of scikit-learn will be technologically sufficient to meet our client's needs. If classification of data requires deep learning to train our classifier, EMELIA, then the team will have to explore alternatives to reach our goals.

One possible library that we could utilize is Pandas, which is a popular open-source library that extends the functionality of Python to allow for data manipulation and analysis. Pandas is an easy to use library that contains tools for reading and writing data between data structures and different formats including CSV, text files, Microsoft Excel Spreadsheets, and SQL databases. While Pandas works well with other python packages such as scikit-learn, it might not be suitable for our machine learning classifier so a framework such as TensorFlow would be more suitable.

We as a team agreed that the utilization of Tensorflow is our best option for success with this project. The capabilities of handling large computations and the specific libraries made for

implementing neural networks was the deciding factor for the use of this framework. In addition, Tensorflow's compatibility with the python language and high-level APIs that support this framework such as Keras make this framework the most optimal when compared to the others.

3.3 Version Control

To maintain quality and efficiency we need to utilize version control software. This also ensures that the contributions are tracked and allows for collaboration on this project. We are looking for software that can meet the minimum requirements for accessibility, capabilities, and integration.

Technologies we evaluated for version control:

- GitHub
- GitLab
- PerForce

3.3.1 GitHub

This is the most popular hosting service for command line repository manipulation. GitHub provides a web-based GUI and multiple features involving collaboration such as local branching, and workflow management. GitHub will be analyzed based on the following criteria: cost, familiarity, usability, and features.

- Cost - Github is free software that can be utilized for both personal and open-source projects.

- Familiarity - The entire development team has utilized GitHub in previous projects, so no additional training is required, allowing easy integration.

- Usability - Github contains the minimum requirements we need for it to be integrated.

- Features - Github contains a variety of features that allow for the use and management of software.

3.3.2 GitLab

GitLab also seemed to be a viable option being that it offers the minimum requirements we were looking for in a version control software. However, we also thought GitLab offered features that were irrelevant to our project expectations and think this can cause added confusion in the future. GitLab will be analyzed based on the following criteria: cost, familiarity, usability, and features.

- Cost - GitLab is not free software and costs $19 per user per month.
- Familiarity - Not a single member has used GitLab, it is not a tool we are familiar with.
- Usability - GitLab contains the minimum requirements we need for it to be integrated.
- Features - GitLab contains a variety of features that allow for the use and management of software such as issue tracking, and CI/CD.

3.3.3 Perforce

Perforce is an enterprise version management system that allows users to connect to a shared file repository. Perforce applications are used to transfer files from the main repository to user workstations. Perforce utilizes a command-line interface, visual client, or web client to make changes to file repositories. This version management system is similar to GitHub since each working repository must be saved on a local hard drive. Perforce will be analyzed based on the following criteria: cost, familiarity, usability, and features.

- Cost - Perforce is not free software and costs a substantial amount for a license.
- Familiarity - Not a single member has used Perforce.
- Usability - Perforce contains the minimum requirements we need for it to be integrated.
- Features - Perforce contain a variety of features that allow for the use and management of software.

Table 4

	Usability	**Cost**	**Familiarity**	**Features**	**Choice**
GitHub	✔	✔	✔	✔	✔
GitLab	✔	✖	✖	✔	✖
Perforce	✔	✖	✖	✔	✖

Conclusion

Our team has looked into three methods of version control for this project and decided to utilize GitHub as our centralized service for our team repositories. With respect to usability and features, all of the following are very similar and offer great services, however, GitHub seems to be the simplest option being that everyone has experience with this software. In addition, GitHub is the most cost-efficient option for our team, offering a free service compared to monthly subscriptions that GitLab and Perforce offer.

Since every member of this team has had some experience with GitHub, and GitHub meets our requirements for what we are looking for in a version control software, we decided that GitHub is the best choice of version control software for the development team. We are confident in this choice and have no doubt that GitHub will strongly contribute to the success of our project.

3.4 Progress Management

Tracking our issues and tasks is a major component of our project's success. With upcoming tasks and requirements, we needed to come up with a solution so that we can work through these challenges effectively. Our approach is to find software that can overall assist our team in project management. Some software we are considering are:

3.4.1 Trello

Trello is a web and mobile application that organizes tasks in a Kanban-style list format. Trello allows the team to organize the list of tasks into different categories based on the current completion status, such as To Do, In Progress, and Done. Trello makes it easy to move individual items to different categories as they are completed. Trello allows users to integrate the different apps that the team utilizes directly into the workflow and syncs across different devices, improving collaboration between team members. Trello's flexibility allows it to be used for both personal and business and contains features that suit most work environments. [10]

Trello will be ranked using the following categories: cost, familiarity, usability, and features.

- Cost - Trello is free software that can be utilized for both personal and business projects.
- Familiarity - The entire development team has utilized trello at least once, so no additional training is required, allowing easy integration into our workflow.
- Usability - Trello has a simple drag-and-drop interface that allows users to seamlessly move tasks from one category into another.
- Features - Trello contains a variety of features that allow for use in any team's work style.

3.4.2 Asana

Asana is a free web and mobile application that allows teams to organize and track their work in real-time. Asana monitors the status of projects and the workloads of each team member to ensure that no one is overworked. Asana also allows the team to work with a calendar in order to keep track of important deadlines, create tasks, assign tasks, and mark tasks as completed. Using color coordination, Asana allows teams to organize tasks based on color logic to prioritize the completion of certain tasks. Our current team has not utilized Asana in previous projects, so we would have to conduct additional training to be more familiar with Asana. [11]

Asana will be ranked using the following categories: cost, familiarity, usability, and features.

- Cost - Similar to Trello, Asana is a web and mobile application that will not cost the development team anything as it is free to use.

- Familiarity - No members of our team have used Asana in previous projects, so some amount of training will be necessary if we are to integrate Asana into our current workflow.

- Usability - Asana utilizes a calendar to keep track of tasks in real-time.

- Features - Asana contains a multitude of features that allows teams to keep track of important deadlines, create tasks, assign tasks, mark tasks as completed, and prioritize the completion of certain tasks using color coordination.

3.4.3 Smartsheet

Smartsheet is a software that provides a service for collaboration and progress management. Smartsheet allows the users to delegate tasks, track progress, manage calendars and share documents. Smartsheet does not offer a free service whereas the cost for this service is $14 a month for individuals and $25 per month for the business plan. In terms of familiarity, our team has not used Smartsheet previously, however, the interface is in tabular form which makes the content organized and the service more user-friendly.
Smartsheet will be ranked using the following categories: cost, familiarity, usability, and features. [12]

- Cost - Smartsheet is not free to use, unlike Trello and Asana, and requires the development team to purchase an individual or business plan that will have recurring charges every month.

- Familiarity - Similar to Asana, no members in our team have used Smartsheet in previous projects but we have worked with Google Drive so some training is required to integrate into our current workflow.

- Usability - Smartsheet integrates with Google and can import templates that the development team can utilize to keep the formatting the same.

- Features - Smartsheet allows the development team to easily assign tasks, track progress completion status, manage calendars between team members, share documents and the ability to have any number of collaborators on a single project.

Table 5

	Cost	Familiarity	Usability	Features	Choice
Trello	✔	✔	✔	✔	✔
Asana	✔	✖	✔	✔	✖
Smartsheet	✖	✖	✔	✔	✖

Conclusion
Trello is a simple web application that organizes tasks in a kanban style format. Trello makes it easy to organize tasks based on completion status and move around tasks as they are worked on by the team. Trello has been used by every team member, so integration into our current workflow will be simple and easy to manage. Asana is a web and a mobile-based software application that organizes available tasks in real-time to ensure that they are worked on and tracks each team member's contribution. By tracking the contributions of each team member, Asana ensures that no one is overworked or underworked. While Asana is a useful tool to keep track of the assigned task for each individual team member, the team has never used it so there would need to be training for all members to integrate it into our current workflow. Smartsheet is a cloud-based platform for businesses to increase the effectiveness of work production by allowing teams to collaborate during project management. Smartsheet allows users to directly attach any type of files directly to a project and can be managed using Google Drive. Although Smartsheet is the most sophisticated software compared to the other software, the cost is something that this team is not willing to pay for this service.

3.5 Language

This section will make an analysis of each language and summarize the metrics in a table. As stated previously, choosing a language deficient in terms of available libraries or lack of integration with a machine learning framework, may cause the implementation of this project to be technologically unfeasible. There are three languages that will be analyzed in this section: Python, JavaScript, C++.

3.5.1 Python

Python is a common programming language currently used in modern data science, informatics, and machine learning projects. Since Python is widely used, there is a large community and

abundance of resources for this language. Currently, Python is the preferred language for machine learning projects such as product recommendations shopping on Amazon and Gmail which is made by Google. Regarding the metrics mentioned in the Technological Challenges (Section 2.5), Python is a high priority. [5]

- Usability
    - Python is a language that is considered easy to write by most programmers. It requires fewer lines of code to perform the same operations as other programming languages.
- Features
    - Python has an expansive list of available libraries and tools to perform the data computation required by a machine learning classification model. Two major data science libraries available to Python include Pandas and scikit-learn.
- Framework Integration
    - Python integrates well with frameworks such as TensorFlow and the SciPy ecosystem. Python is the primary language used for machine learning tools that organize data structures and perform high computing for these machine learning frameworks.
- Documentation
    - Python is a widely used language with excellent documentation and online support. The documentation extends from introductory tutorials through the use of complex computing libraries such as Pandas or scikit-learn.

3.5.2 JavaScript

JavaScript is an interpreted language that is mainly used for web development. This language is mostly used to build front-end and back-end systems for websites, but it has some applications for machine learning. This language has an application for machine learning and is continuing to grow in use within the field.

- Usability
    - This programming language is easy to use and builds off of the syntax that is used from Java.
- Features
    - JavaScript has a few libraries that are available for machine learning. More specifically, this language seems to be at a deficit in comparison to other programming languages when it comes to available tools.
- Framework Integration

    - JavaScript integrates well with TensorFlow.js. DialogFlow is another powerful tool that is neither a framework or a library, but a technology developed by Google for AI development.

- Documentation
    - The documentation for JavaScript compares well with other languages. Since the presence of this language is low in machine learning related projects, the development team might find it difficult to utilize the language as needed during implementation.

3.5.3 Java

Java is a strongly typed language with performance benefits over other popular machine learning languages like Python. Java is the programming language that the development team has the most experience using for project development. [5]

- Usability
    - The required learning for this language should only pertain to the use of machine learning. All members of the development team should be able to utilize the language to structure classes and relate each class accordingly.
- Features
    - There are a few major libraries that assist with deep learning models and computational tools for this language. The uses of these major libraries are typically integrated to accomplish major machine learning tasks.
- Framework Integration
    - TensorFlow supports an API for Java. Java seems to have lower integration than other languages for available machine learning frameworks and tools.
- Documentation
    - Overall, Java is a well-documented language with plenty of resources that are available online. It is a language that has been used for a large period of time in the industry and has a large community of users.

Table 6

	Usability	Features	Framework Integration	Documentation	Choice
Python	✔	✔	✔	✔	✔
JavaScript	✔	✖	✔	✔	✖
Java	✔	✖	✖	✔	✖

Conclusion

For the purpose of developing EMELIA and creating an accurate intelligent assessor, Python is the programming language that would best allow us to reach our goals. Overall, Python is a ready-to-use language that greatly supports the various aspects of this project. Developing EMELIA utilizing this language seems practical from a technological standpoint. It is highly unlikely that the development team would suffer from technical issues, issues learning the language or completion of the project using Python as a programming language. It has low overhead for learning required by the development team. Python has a large feature set of computational and deep learning libraries. TensorFlow integrates well with Python as a machine learning framework. There is comprehensive documentation to support development for this project, which makes it ideal for the development of this project. [9]

4. Technological Integration

This section will provide a conclusion to our technological analysis section. We will define which tools from the analysis will integrate cohesively to build EMELIA. Each of the feasibility challenge tools mentioned previously in the document will be examined, rated, and compared. The final selection for these tools will be based on cost, usability, features, and documentation.

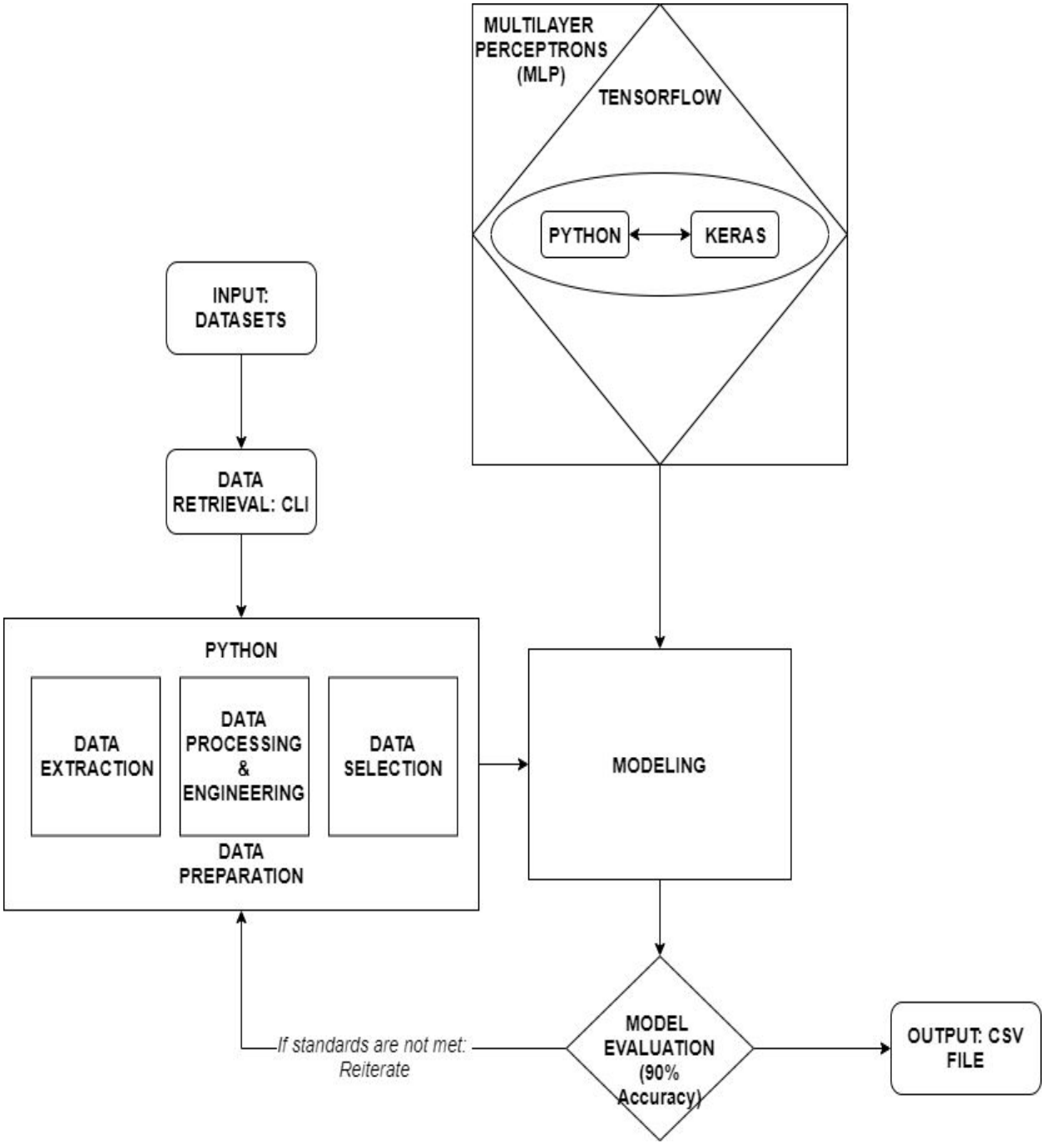

MULTILAYER PERCEPTRONS (MLP)
TENSORFLOW
PYTHON
KERAS
INPUT: DATASETS
DATA RETRIEVAL: CLI
PYTHON
DATA EXTRACTION
DATA PROCESSING & ENGINEERING
DATA SELECTION
DATA PREPARATION
MODELING
MODEL EVALUATION (90% Accuracy)
If standards are not met: Reiterate
OUTPUT: CSV FILE

Figure 2

Figure 2 depicts how we envision we are going to approach our problems with the technologies we have chosen. Below is an overview of what the diagram represents in more detail.

1. Our system will need to be able to take in input in the format of a CSV file representing several “tickets”. We will then retrieve this data by implementing a program to read in the arguments through the command-line interface. We were instructed that the client prefers this method and no graphical user interface is necessary.

2. These tickets will be similar in structure but may have differences by the number of classifications or categories within each ticket. However, there are a few categories that are consistent throughout the series tickets such as a date, or region. We will create a program using Python that will parse through each ticket and extract this data and store it for the machine learning model. This is the data preparation step and is vital for the next step “modeling”.

3. The modeling step in figure 2 is the collection of data being analyzed using our machine learning algorithm that incorporates the classifying model, TensorFlow as the framework, and the implementation of the language Python using Keras as the supporting library. Our system will then have to make decisions on where to categorize each ticket based on the learning algorithm that will analyze similarities and patterns.

4. The program will then proceed to test the results as shown in the model evaluation step and in the event the results does not meet our goal of a minimum accuracy of 90%, the program reverts back to the data preparation/extraction step for further evaluation and repeats the following steps until a minimum accuracy of 90% is achieved.

5. Assuming the program had successfully progressed through all the steps explained in the diagram, the results are ready to be returned back to the user. We will integrate a solution in which the results of the category each ticket is classified in will be expressed in another column within the same CSV file alongside each ticket. After allocating memory for each ticket’s category result we will then update the CSV file where the user can access these results.

Given the overview of the diagram, additional details concerning the specific technologies and their roles in this project will be explained below.

With the many different neural networks that exist, our team proceeded to undergo extensive research to find the best modeling system. We decided to use a multilayer perceptron modeling system for our overall approach because of its modularity as a machine learning modeling and

the potential it has on the sophistication of this product. We then encapsulate a framework to work within this modeling system that will be used as a building block and provide functionality for our machine learning. We went with Tensorflow because of its documentation and superiority over other machine learning frameworks. Being that Tensorflow is the most widely used framework for machine learning, it is also compatible with the most popular language used for machine learning, Python. Not only that, Python is considerably easy to work with and has an expansive list of available libraries and tools. We as a team think this is the best language to be implemented with our framework and expect little to no problems when integrating the two technologies. Not only will Python be used to code our machine learning, but will also be utilized to implement the data extraction.

For better assistance we will also be working with other open-source neural network libraries such as Keras. Keras is also written in python and works well with Tensorflow thus will be a great tool to help us engineer the project features and complex methods.

5. Conclusion

The objective of our project is to create an interactive machine learning intelligent assessor for General Dynamics. With the help of our sponsors, Rick Duarte and Jon Lewis, we will develop a system that allows for the retrieval, analysis, and exportation of data that will assist the organization in failure reporting.

Our envisioned solution will utilize the Multilayer Perceptron learning model to utilize data as input and provide an accurate classification of system failure data. The TensorFlow machine learning framework will allow the development team to use our selected learning model. Due to the popularity and community support surrounding this machine learning framework, the development team is confident this framework will benefit the development of EMELIA. Github is the version control system to be used by the team. Github is a widely used version control system that enhances development throughout various stages of implementation. Trello will be used to track issues and task management for this project. The development team will utilize an Agile development method with Trello as a Kanban board. This specific method will aid the team in meeting deadlines and provide support for team members as they run into obstacles during development. Please see the table below for a summary of proposed solutions:

Table 8

Technological Challenge	**Proposed Solution**
Model Classification	Multilayer Perceptrons
Framework	Tensorflow
Version Control	GitHub Software
Project Management Approach	Trello Software
Language	Python

Overall, the choice of technologies will make our envisioned solution both technologically feasible and less complex to implement. Our technological choices will provide the functionality needed to progress EMELIA into a finished product. Based on the rationale stated throughout the subsections of this document, the development team is confident in our ability to build our Event-driven MachinE Learning, Intelligent Assessor (EMELIA).

6. Resources

[1] Robert Bradshaw and Stefan Behnel. 2007. C-Extensions for Python. (July 2007). Retrieved November 4, 2019 from https://cython.org/

[2] Radoslaw Fabisiak. 2019. The best programming language for Artificial Intelligence and Machine Learning. (August 2019). Retrieved November 4, 2019 from https://medium.com/duomly-blockchain-online-courses/the-best-programming-language-for-artificial-intelligence-and-machine-learning-538486b462c

[3] Fernando Perez. 2001. Jupyter and the future of IPython. (2001). Retrieved November 4, 2019 from http://ipython.org/

[4] Travis Oliphant, Pearu Peterson, and Eric Jones. 2001ciPy.org. (2001). Retrieved November 4, 2019 from https://www.scipy.org/

[5] Fernando Perez. 2001. Overview and getting started. (2001). Retrieved November 4, 2019 from https://ipyparallel.readthedocs.io/en/latest/intro.html

[6] Fernando Perez. 2015. Project Jupyter. (2015). Retrieved November 4, 2019 from https://jupyter.org/

[7] Wes McKinney. 2008. Python Data Analysis Library. (January 2008). Retrieved November 4, 2019 from https://pandas.pydata.org/

[8] David Cournapeau. 2007. scikit-learn Machine Learning in Python. (2007). Retrieved November 4, 2019 from https://scikit-learn.org/stable/

[9] Google Inc. 2015. An end-to-end open source machine learning platform (2015). Retrieved November 4, 2019 from https://www.tensorflow.org/

[10] Finnegan, M. 2018. What is Trello? A guide to Atlassian's collaboration and work management tool. (2018). Retrieved November 6, 2019 from https://www.computerworld.com/article/3226447/what-is-trello-a-guide-to-atlassians-collaboration-and-work-management-tool.html.

[11] Moore, B. and Duffy, J. 2018. Asana Review & Rating. (2018). Retrieved November 6, 2019 from https://www.pcmag.com/review/301098/asana.

[12] Finnegan, M. 2017. What is Smartsheet? A spreadsheet-based project management tool. (2017). Retrieved November 6, 2019 from https://www.computerworld.com/article/3239150/whats-smartsheet-a-spreadsheet-based-tool-for-simpler-project-management.html.

[13] Bronshtein, A. 2017. A Quick Introduction to the "Pandas" Python Library. (2017). Retrieved November 6, 2019 from https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library-f1b678f34673

[14]Sakshi Indolia. 2018. Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. *Procedia Computer Science* (2018).

[15]Tomas Mikolov. 2012. Context dependent recurrent neural network language model. *Institute of Electrical and Electronics Engineers* (December 2012).

[16]Jason Brownlee. 2019. When to Use MLP, CNN, and RNN Neural Networks. (August 2019). Retrieved November 8, 2019 from https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/