National College of Ireland

Ensemble Machine learning to Detect Exoplanets.

Configuration Manual

MSc Research Project
Data Analysis

Vishal Petkar
Student ID: x21216461

School of Computing
National College of Ireland

Supervisor: Cristina Hava Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. ……Vishal Petkar……

Student ID: ……x21216461…….……

Programme: ……Data Analysis…… **Year:** …2023-2024

Module: ……Research Project…….……

Lecturer: …… Cristina Hava Muntean …….……

Submission Due Date: ……25/04/2024…….……

Project Title: Ensemble Machine learning to Detect Exoplanets

Word Count: ……4…… **Page Count:** ……..……

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: …Vishal Petkar……

Date: ……24/04/2024……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

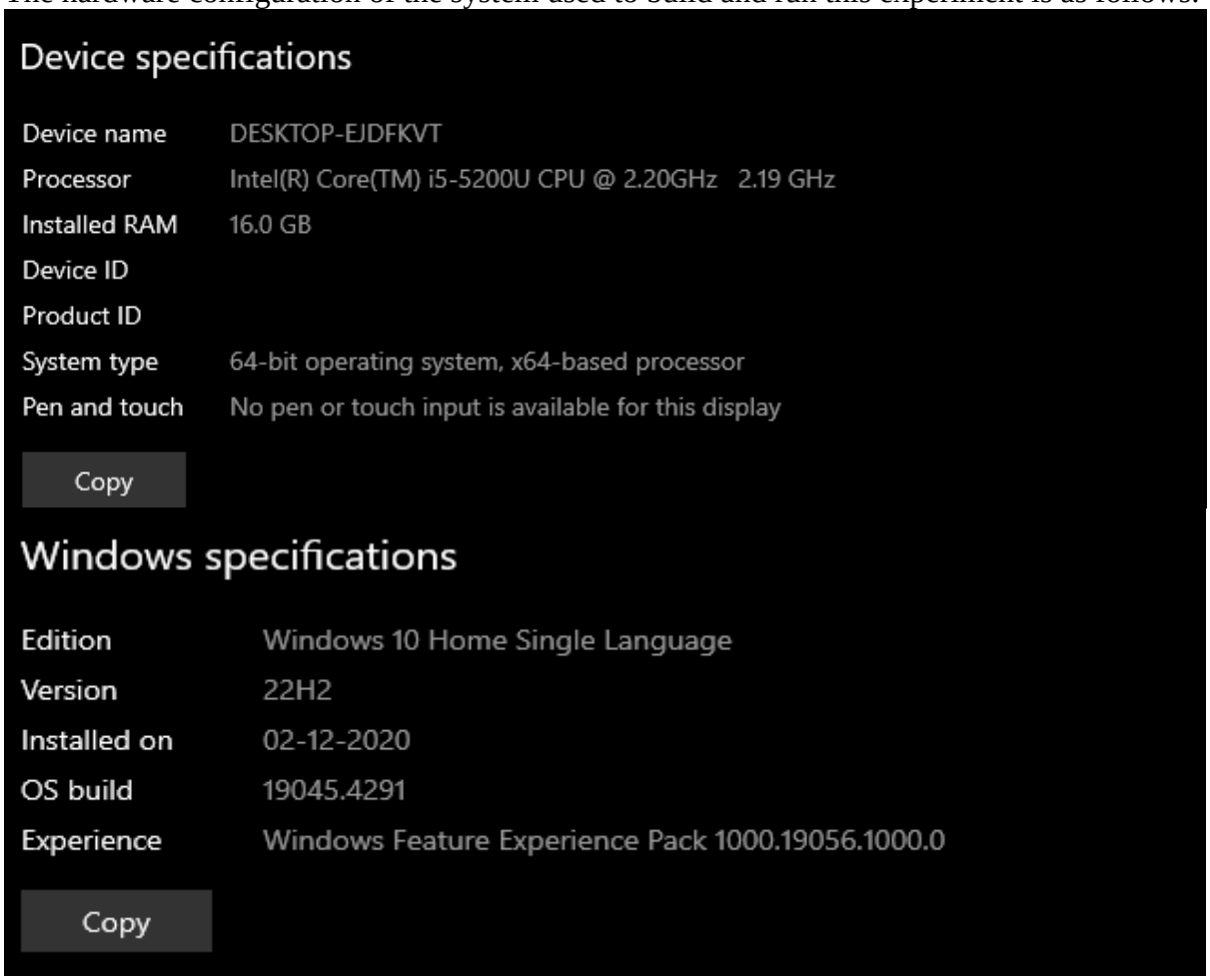
Vishal Petkar
Student ID: x21216461

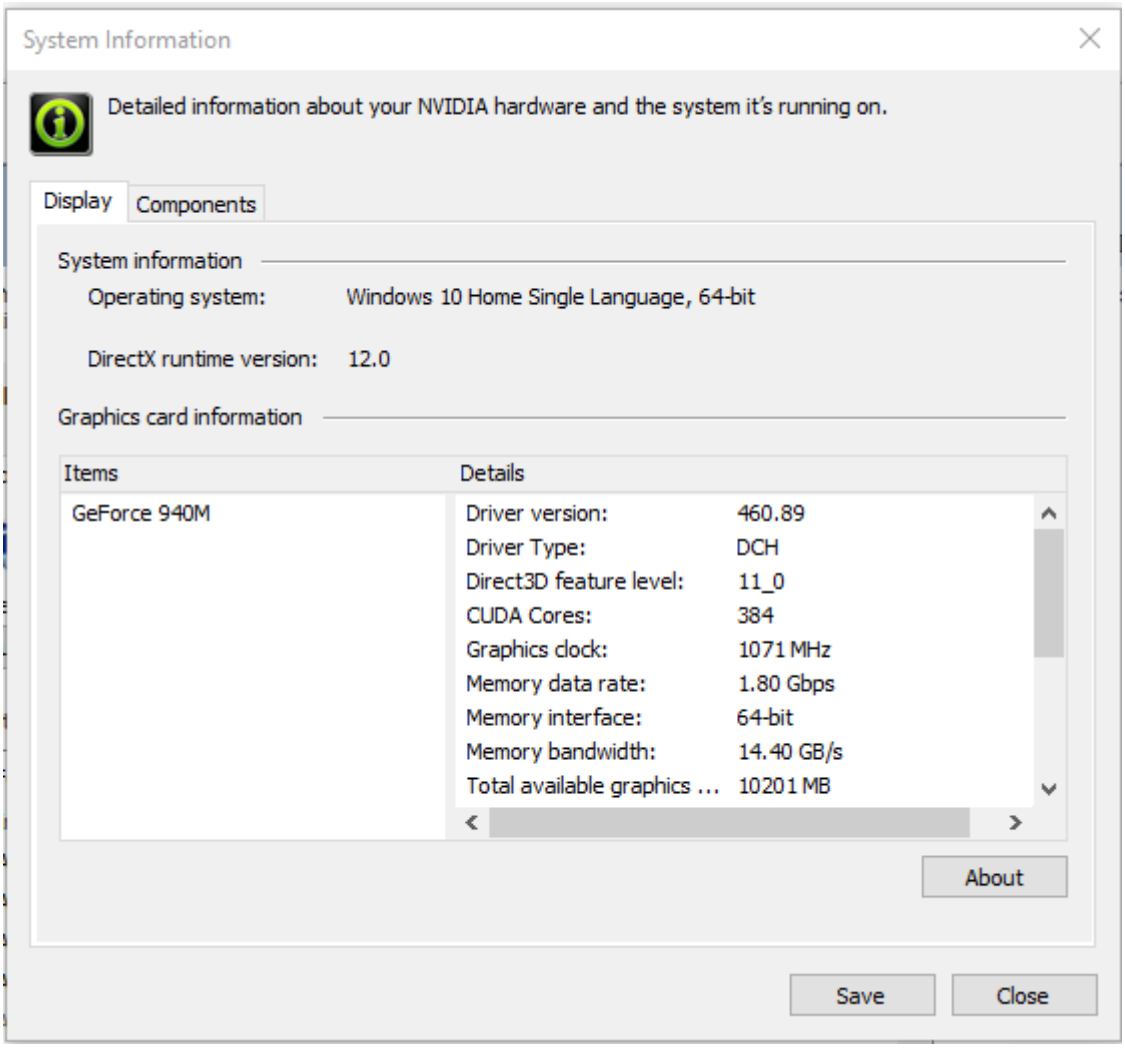
1 Introduction

This document contains the detailed instruction on how to replicate the experiment from the main research paper. The configuration manual discusses the machine requirement needed to build and run this model. The steps to all the installation required are mentioned in this document.
This experiment requires python version 3.9.16 to be installed in the system.

2 Hardware configuration

The hardware configuration of the system used to build and run this experiment is as follows:



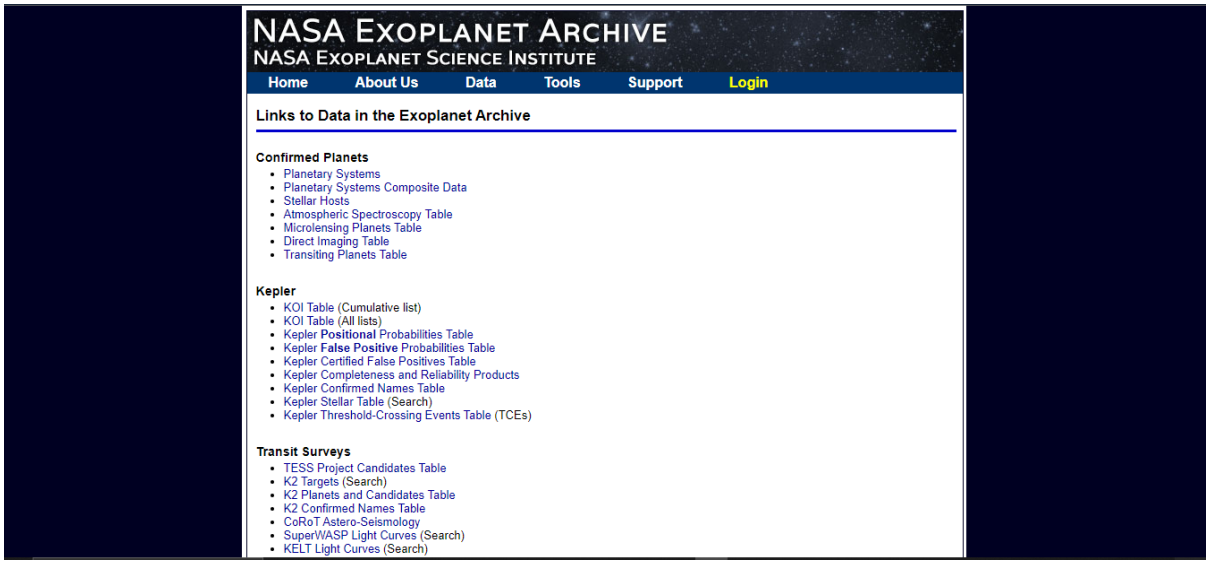


3 Project Files

This section describes the project files needed to replicate the experiment.

Pre-requisite:
The system should have lightkurve python module installed and associated visualization modules such as matplotlib, seaborn etc. The CNN model was created using TensorFlow and the system available Nvidia GPU. If GPU is not present in system, the file can be run on Google collab for faster execution.

Dataset:
The dataset containing the confirmed exoplanets data and confirmed false positive data was obtained from https://exoplanetarchive.ipac.caltech.edu/



Once the data is downloaded, the necessary data such as the TOI or KOI star names and the orbital period data is extracted and put in another excel sheet for easy access. This data is then read by the jupyter file and the associated lightcurves of the TOI/KOI stars are generated and saved.

4 Software used:

- Microsoft Excel for maintain initial dataset.
- Jupyter Notebook for coding the model and evaluation.
- TensorFlow for using the available system Nvidia GPU for training the CNN model.

5 Replicating the experiment:

- To generate the lightcurves, import the lightkurve module and download the lightcurve data using the search_lightkurve command.

```
from lightkurve import search_targetpixelfile
from lightkurve import TessTargetPixelFile
import matplotlib.pyplot as plt
%matplotlib inline
import lightkurve as lk
import numpy as np
import pandas as pd
import os

# Read the Excel file
stars_df = pd.read_excel(r'Thesis files\LightCurve_images\FP_names.xlsx')

# Extract star names and corresponding orbital periods
star_names = stars_df['TESS'].tolist()

orbital_periods = stars_df['TESS_orbit'].tolist()

# Create directories to save plots if they don't exist
folded_dir = r'Thesis files\LightCurve_images\TESS_orbit_FP_folded'
binned_dir = r'Thesis files\LightCurve_images\TESS_orbit_FP_binned'
os.makedirs(folded_dir, exist_ok=True)
os.makedirs(binned_dir, exist_ok=True)

for star_name, period in zip(star_names, orbital_periods):
    try:
        # Search and download light curve for each star
#         print(star_name)
        lc = lk.search_lightcurve('TIC ' + str(int(star_name)) , mission='TESS' , cadence='long').download()
```

Once all the lightcurves are generated, they are stored in separate files for training and testing. The files are then read in by the models that are trained. Once the models are trained, they are loaded into a script to turn it into an ensemble model.

```
import os
import cv2
import joblib
import numpy as np
import pandas as pd
import tensorflow as tf
from tensorflow.keras.models import load_model
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score, f1_score
import seaborn as sns
import matplotlib.pyplot as plt


# Load models
random_forest_model = joblib.load("models/random_forest_Training_70_30.pkl")
svm_model = joblib.load("models/svm_binned_training_70_30.pkl")
keras_model = load_model("models/CNN_Training_70_30.h5")
```

The files need to be executed in the following order for the project to work

1. Lightkurve_generation_file.ipynb
2. Kepler_dataAnalysis.ipynb
3. CNN.ipynb
4. Random Forest.ipynb
5. KNN.ipynb
6. SVM_lightcurve.ipynb
7. Ensemble_lightcurve_code.ipynb

The lightcurve data files that is used for training and testing and the created models (CNN, KNN, SVM and Random forest) are saved on my google drive and can be accessed with the following link.
https://drive.google.com/drive/folders/1bHKSrFBcqKJlM1fjqJnrFaDe0S0o9z2M?usp=drive_link