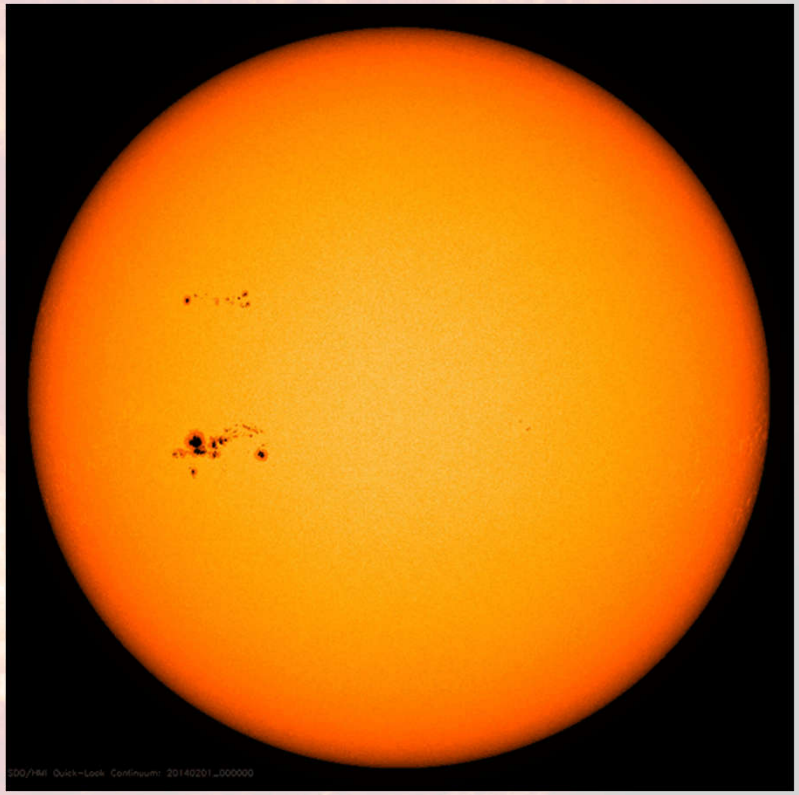
# INTELLIGENT SUNSPOTS DETECTION & FORECASTING USING ADVANCED MACHINE LEARNING

by

Muhammad Ali Ismail

UN / Azerbaijan Workshop on ISWI
2022 October 31 – Nov 4, Baku, Azerbaijan



The NED University of Engineering & Technology is a public university located in the urban area of Karachi, Sindh, Pakistan. It is one of the oldest and best engineering universities in Pakistan, acknowledged for its best teaching practices and graduates.

**The National Center in Big Data and Cloud Computing (NCBC) focuses on R&D and human resource development in the specialized field of Big Data and Cloud Computing and its practical applications, which are important components of Vision 2025. The role of Big Data Analytics and Cloud Computing is growing in many businesses and applications domains and has become extremely critical to economic growth and national competitiveness. NCBC aims at becoming the leading hub of innovation, scientific research, knowledge transfer to local economy, and training in the area of data analytics, cloud computing, and data science.**

NATIONAL CENTER IN BIG DATA AND CLOUD COMPUTING

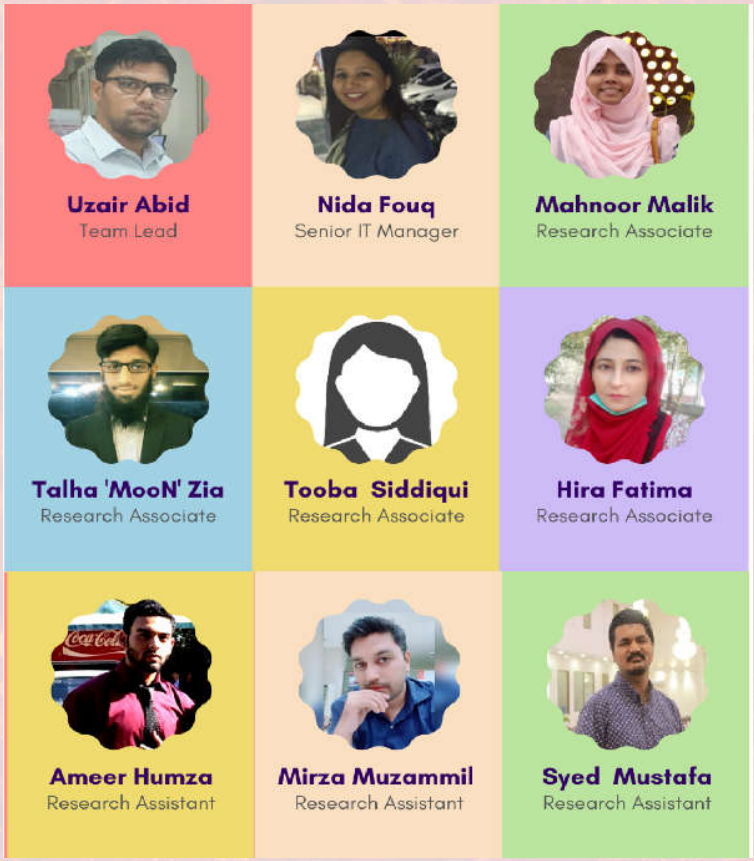
# RESEARCH TEAM

## OUR TEAM

**Dr Muhammad Ali Ismail**

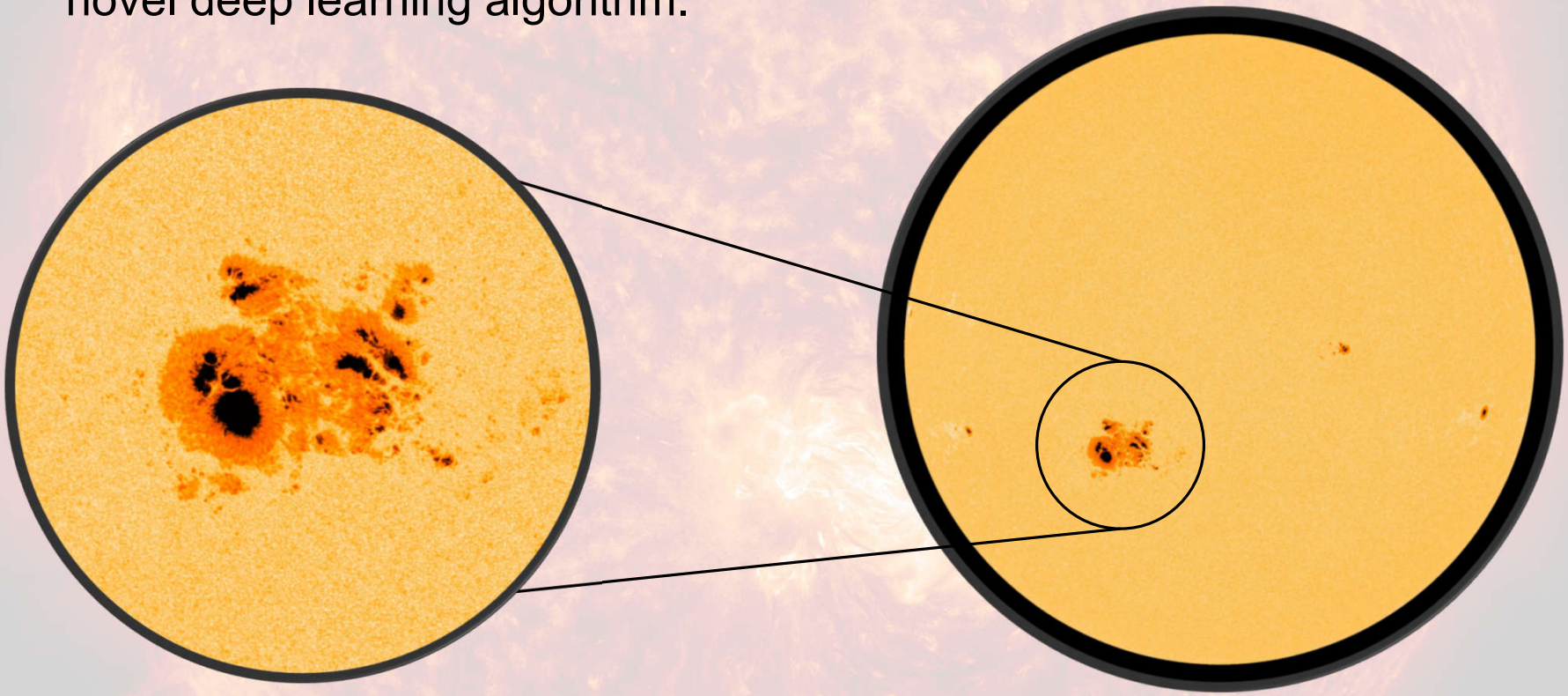
*Principal Investigator*

| maismail@neduet.edu.pk |



# PROJECT IDEA

- To automatically detect sunspots and forecasting solar cycle 25 using novel deep learning algorithm.



# AIMS, OBJECTIVES & SCOPE

- To build **an Open-Data repository containing processed and pre-processed** solar images that researchers across the country can explore for further analysis.
- **To aid researchers and students** in learning about the advanced complex pattern recognition techniques and to work effectively with the Solar Feature Catalogue.
- To **develop web and mobile application** to generate timely space weather alerts and to create public awareness related to impacts of solar dynamics on Earth and space weather as a whole.
- **Astronomy education & research opportunities**.
- **National & International level collaborations**.

## EARLIER PREDICTIONS OF SC 25

- **Forecasts for Solar Cycle 25 have varied from suggesting a stronger cycle than Solar Cycle 24 to the weakest cycle ever recorded with sunspot numbers ranging from 57-167.**
- (Covas et al., 2019) used spatial-temporal data with neural networks to predict that the upcoming Solar cycle 25 would be the weakest cycle ever recorded with sunspot numbers of 57±17 and total sunspot area of ~700 with a peak around 2022-2023.
- (Upton & Hathaway, 2018) used a flux transport model and predicted that Solar Cycle 25 would be similar in size to Solar Cycle 24 with a 15% uncertainty.
- (Labonville et al., 2019) used a dynamo-based model to forecast the upcoming solar cycle and predicted a maximum sunspot number of 89 +29/-15.
- An international panel co-chaired by **NOAA/NASA** released a preliminary forecast on April 5, 2019 with the consensus predicting that Solar Cycle 25 would be similar in size to Solar Cycle 24 and the minimum and maximum sunspot number forecast was 95 and 130 respectively.
- (Pala & Atıcı, 2019) used two layers of stacked LSTMs and predicted that the upcoming Solar Cycle 25 would have a maximum sunspot number of 167.3 with the peak being reached in 2023.2 ±1.1.
- **The low RMSE values of the model and its forecasts after training on Solar Cycle 23 and Solar Cycle 24 give us confidence that our model may work ahead from the best deep neural network based approach using temporal indices of sunspot numbers for predicting the strength of the upcoming solar cycle.**

# LITERATURE REVIEW

| # | Authors Name | Dataset | Method/Algorithm | Results/Findings |
|---|---|---|---|---|
| 1 | Turmon et al. (2002) | SOHO/MDI magnetogram/continuum | Image Segmentation | Sunspots, faculae detection |
| 2 | Qahwaji & Cloak (2007) | SOHO/MDI magnetogram/continuum | Threshold-based Approach | 90% correct match for sunspots detection |
| 3 | Zharkov et al. (2005) | SOHO/MDI magnetogram/continuum | Edge Detection | 96% correlation with NOAA Observatory, USA |
| 4 | Zharkova et al. (2005) | SOHO/MDI magnetogram/continuum | Watershed Transform | Hybrid region growing and Edge detection techniques are the most efficient |
| 5 | Zhang et al. (2010) | SOHO/MDI magnetogram/continuum | Threshold-based Approach | 73.8% TPR |
| 6 | Verbeeck, Higgins & Colak et al. (2011) | SOHO/MDI magnetogram/continuum & EIT UV images | Performance evaluation of four Feature recognition algorithms | All four of them have pros and cons |
| 7 | Martens et al. (2011) | NASA SDO/HMI | Image Classification | 95% agreement with human classified dataset |
| 8 | Zharkov et al. (2004) | SOHO/MDI magnetogram/continuum | Intensity based method | 98% agreement with catalogues |
| 9 | R du Toit et al. (2020) | SOHO/MDI magnetogram/continuum | OpenCV Contour Trace algorithm | 88% detection accuracy |

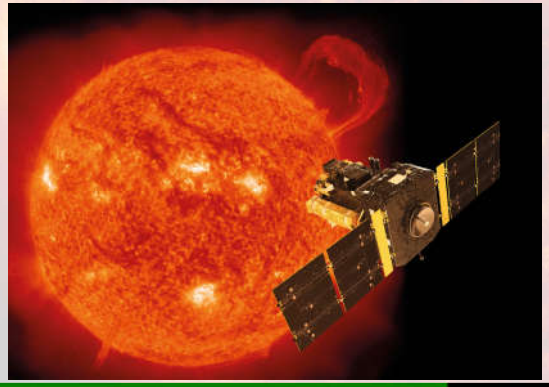
# METHODOLOGY

## DATASETS

### Solar and Heliospheric Observatory (SOHO)

(http://soi.stanford.edu/production/int_gifs.html)

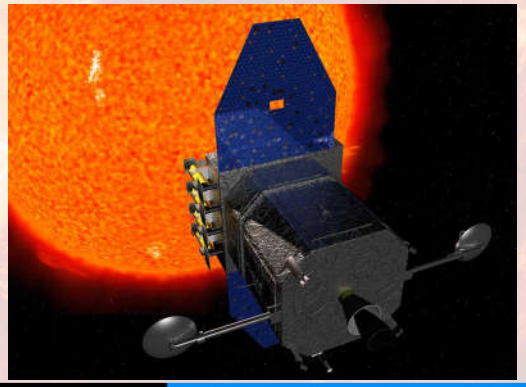
- The Solar and Heliospheric Observatory (SOHO; Domingo et al., 1995) is a joint project between the European Space Agency (ESA) and NASA that was launched on 2nd December 1995. SOHO's studies range from the Sun's interior, its visible surface and stormy atmosphere, to where the solar wind blows in distant regions of our Solar System.



### Solar Dynamics Observatory (SDO)

(http://jsoc.stanford.edu/data/hmi/images/).

- The Solar Dynamics Observatory (SDO; Pesnell et al., 2012) is a NASA mission that was launched on 11th February 2010 to understand the causes of solar variability and its impacts on Earth.



# DATASET STATISTICS

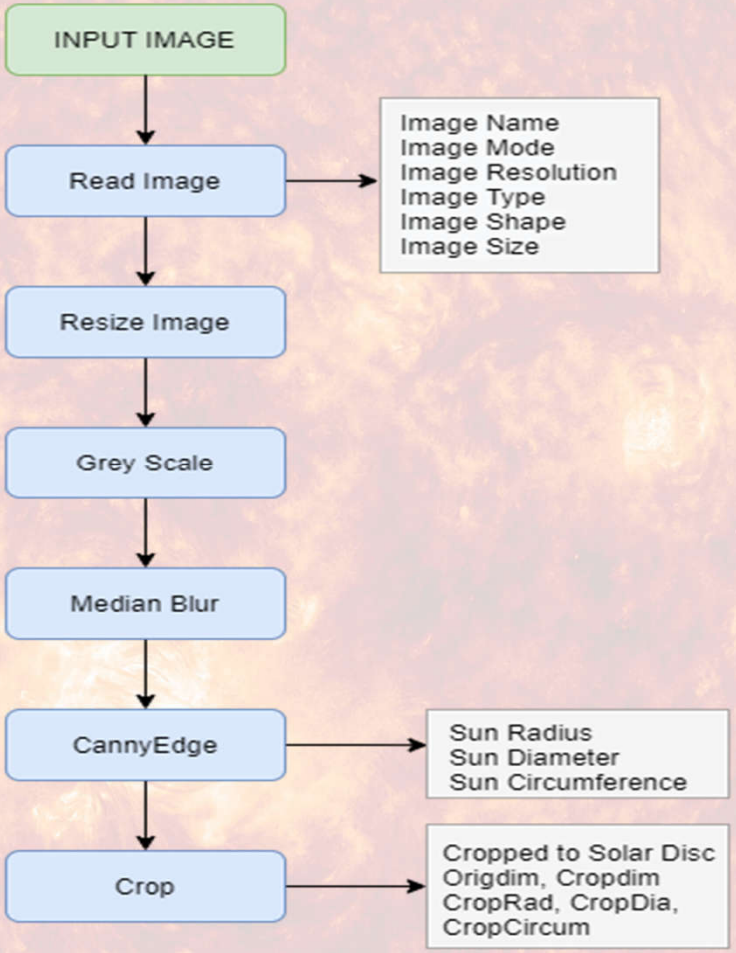
- Two different datasets were used that contain around **2 TB** worth of images.
- Python based source code was written to automatically download the images with respect to date, time and format of files.

| NASA SDO | | NASA SOHO | |
|---|---|---|---|
| FORMAT | JPEG | FORMAT | JPEG |
| SPATIAL RESOLUTION | 4096x4096 | SPATIAL RESOLUTION | 1024x1024 |
| TEMPORAL RESOLUTION | 15 minutes | TEMPORAL RESOLUTION | ~08 hours |
| SIZE | ~5 MB/Image | SIZE | ~0.5 MB/Image |
| DATA DURATION | 2010 - 2020 | DATA DURATION | 1996 - 2009 |
| NO. OF IMAGES | ~350,000 | NO. OF IMAGES | ~14,000 |

# METHODOLOGY

## DATA PRE-PROCESSING

Prior to any automatic detection of sunspots on the images, it is necessary to remove / reduce noise as it hampers any automatic processing algorithm.

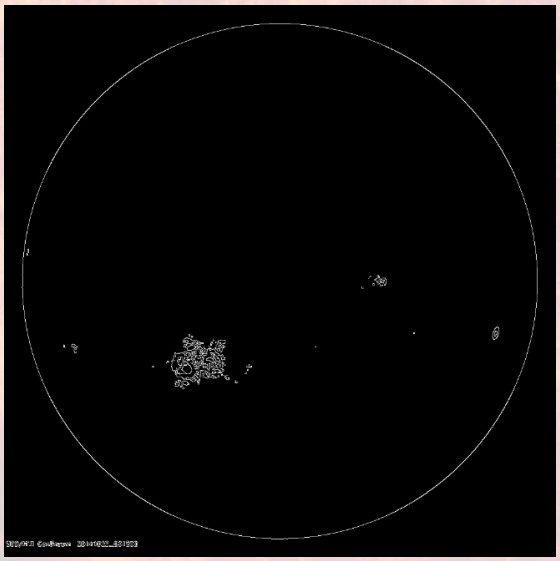


# METHODOLOGY

## DATA PRE-PROCESSING

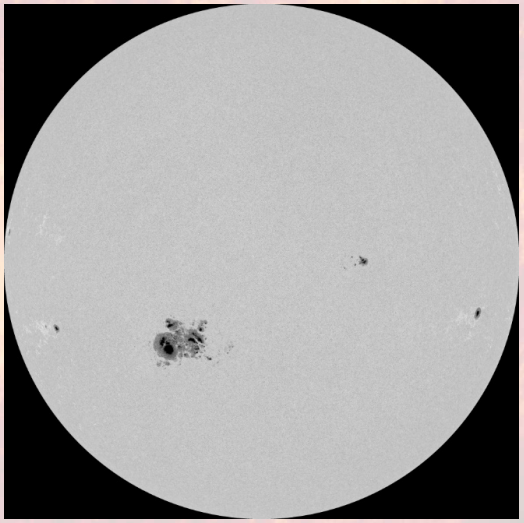
### CANNY EDGE DETECTION

Canny Edge Detection is used here to detect the solar disk and extract the radius and center coordinates of the disk.



### CROPPED TO EDGE

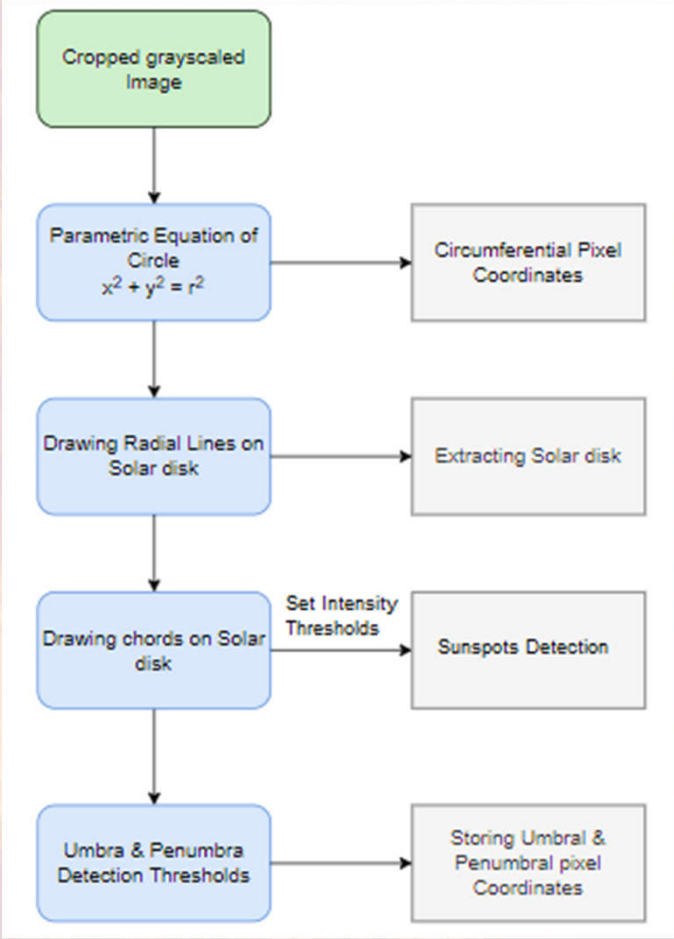
Regions where pixel intensity values were below threshold value were cropped leaving the solar disk in the image.



# METHODOLOGY

## SUNSPOTS DETECTION

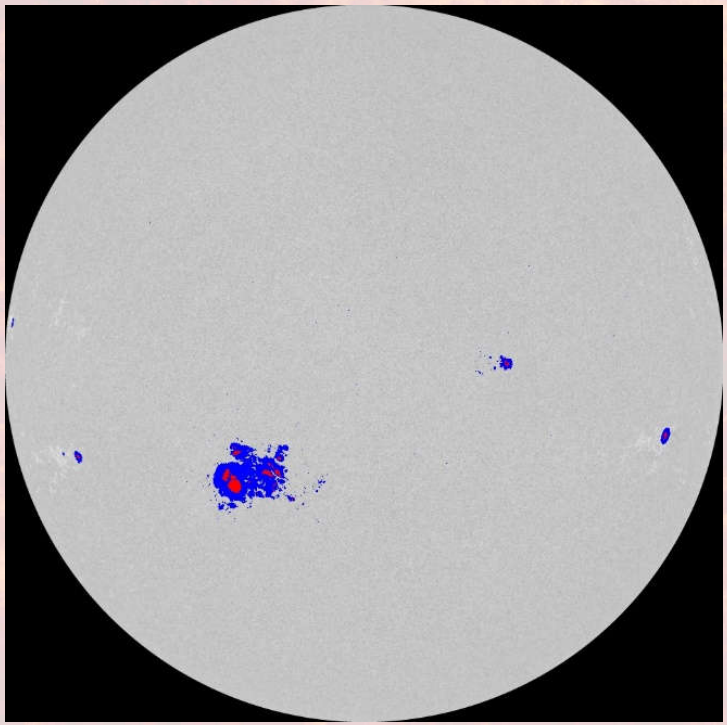
- Pixel coordinates $(x_2, y_2)$ on circumference of solar disk were computed using parametric equation of circle $(x^2 + y^2 = r^2)$.
- Radial lines were then drawn from the center of disk coordinates to edge of solar disk coordinates at **angles ranging from 0˚ to 359˚ incrementing by a factor of 1/9th.**
- Separate thresholds of intensity values were carefully selected to detect **darker umbra and lighter penumbra regions** from the images.
- **Every pixel of the solar disk was read by drawing horizontal and vertical chords** for each quadrant of solar disk and based on the selected threshold values.
- Pixels which have intensity values lower than threshold value were taken into account for each of the umbra and penumbra regions and stored separately in nested lists.



# METHODOLOGY

## SUNSPOTS DETECTION

- The figure below shows detected sunspots with umbra and penumbra regions separately in red and blue colors respectively.

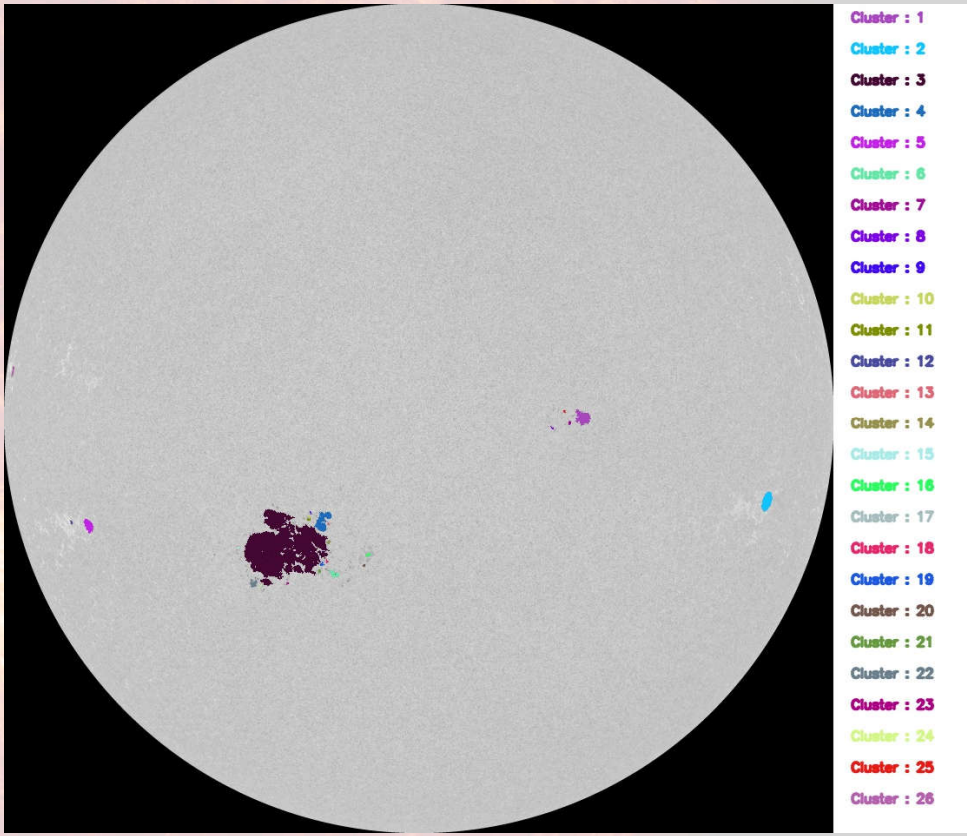


# METHODOLOGY

## CLUSTERING

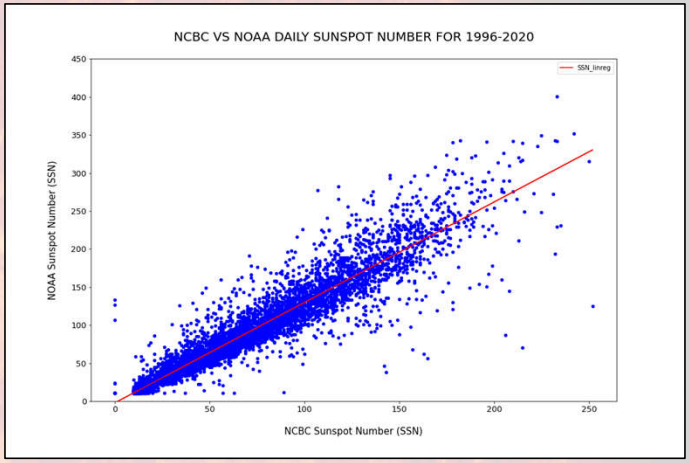
- We developed a clustering algorithm that takes a list of pixel values consisting ‘x’ & ‘y’ coordinates of region of interests and return a list of clusters.

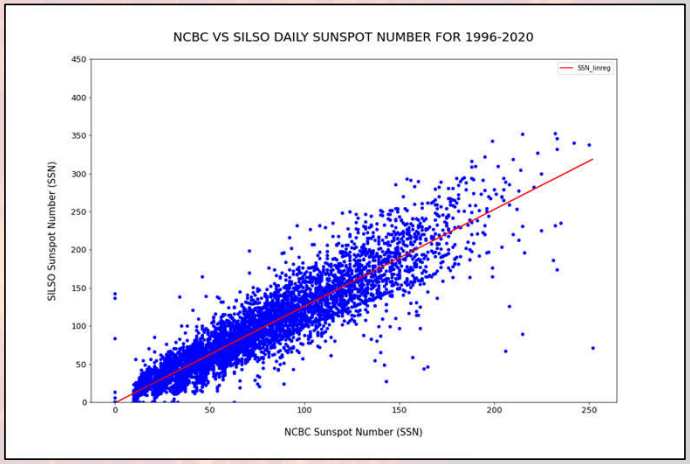
- After detection and clustering following parameters were stored for each month:
  - **Name**
  - **Original Dimensions**
  - **Resized Dimensions**
  - **Canny Radius**
  - **Crop Radius**
  - **Processed Radius**
  - **Number of Umbra**
  - **Umbra Pixel Coordinates**
  - **Number of Penumbra**
  - **Penumbra Pixel Coordinates**
  - **No of Clusters**
  - **Final Cluster Pixel Coordinates**



# VALIDATION DATASETS

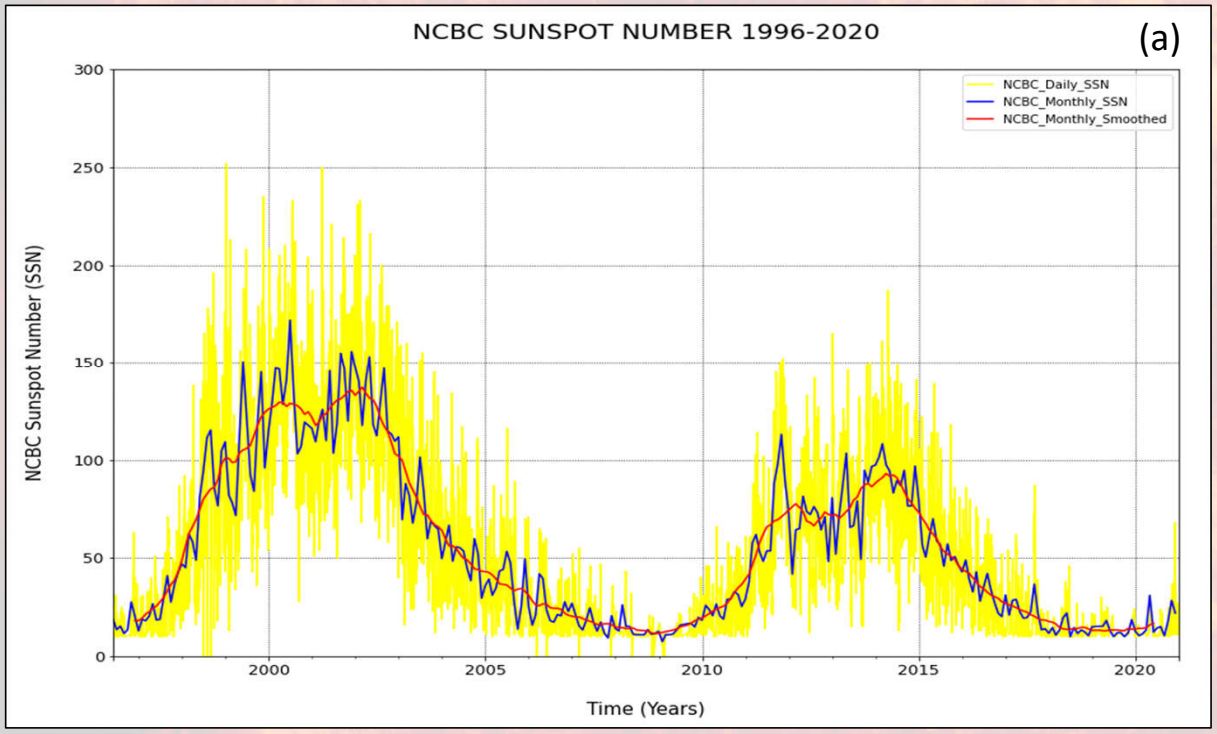
- Two different sunspot number datasets were used to evaluate the performance of algorithm. First dataset is a csv file obtained from the World Data Center **SILSO**, Royal Observatory of Belgium, Brussels (http://sidc.be/silso/datafiles) (SILSO World Data Center, 2021).
- The second dataset is obtained from the website (http://solarcyclescience.com/activeregions.html) made available and maintained by Lisa Upton and David Hathaway. It contains yearly RGO and **USAF/NOAA** data files.
- Data reduction procedures were applied to both SILSO and NOAA datasets prior to perform validation on our dataset to match the parameters present in csv file that was previously cleaned and saved.
- To assess the performance of NCBC dataset, it was correlated with both the SILSO & NOAA datasets.
- Linear regression was then applied on 'NOAA_Daily_SSN' & 'NCBC_Daily_SSN' and 'SILSO_Daily_SSN' & 'NCBC_Daily_SSN' columns as shown in figures (top) & (bottom).

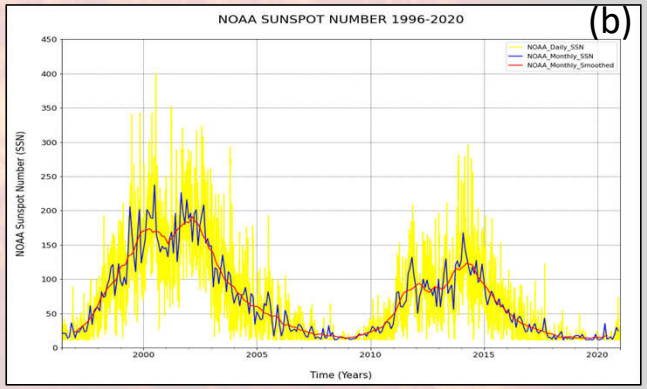


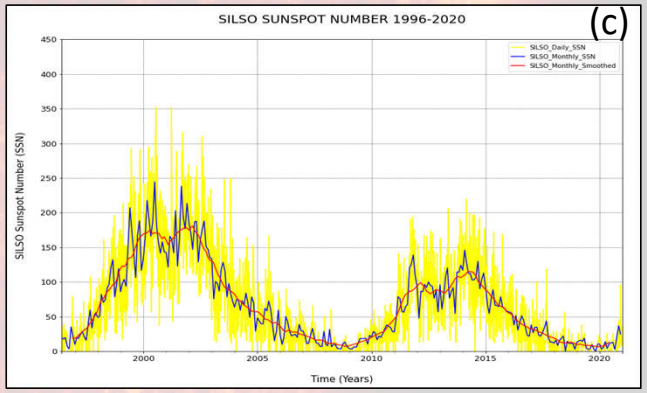



## VALIDATION DATASETS

- The figures (a), (b) & (c) show plots of daily, monthly and smoothed sunspot numbers for both solar cycles 23 & 24 of NCBC, NOAA and SILSO SSN datasets.



# FORECASTING

## N-BEATS (NEURAL BASIS EXPANSION ANALYSIS FOR INTERPRETABLE TIME SERIES FORECASTING):

- N-BEATS (Oreshkin et al., 2020) is a deep neural architecture based on backward and forward residual links and a very deep stack of fully-connected layers.
- Prior to implementing the N-BEATS algorithm the dataset was prepared as a timeseries data with one observation for each day that was smoothed to reduce large variations in the sunspot number values.
- The timeseries was then split into training, validation and test sets. The data compositions for the dataset are given in Table below:

| Series | Sample Size (months) | Training set (months) | Test set (months) |
|---|---|---|---|
| NCBC_Monthly_Smoothed | 295 | 283 | 12 |

**Model Parameters:**

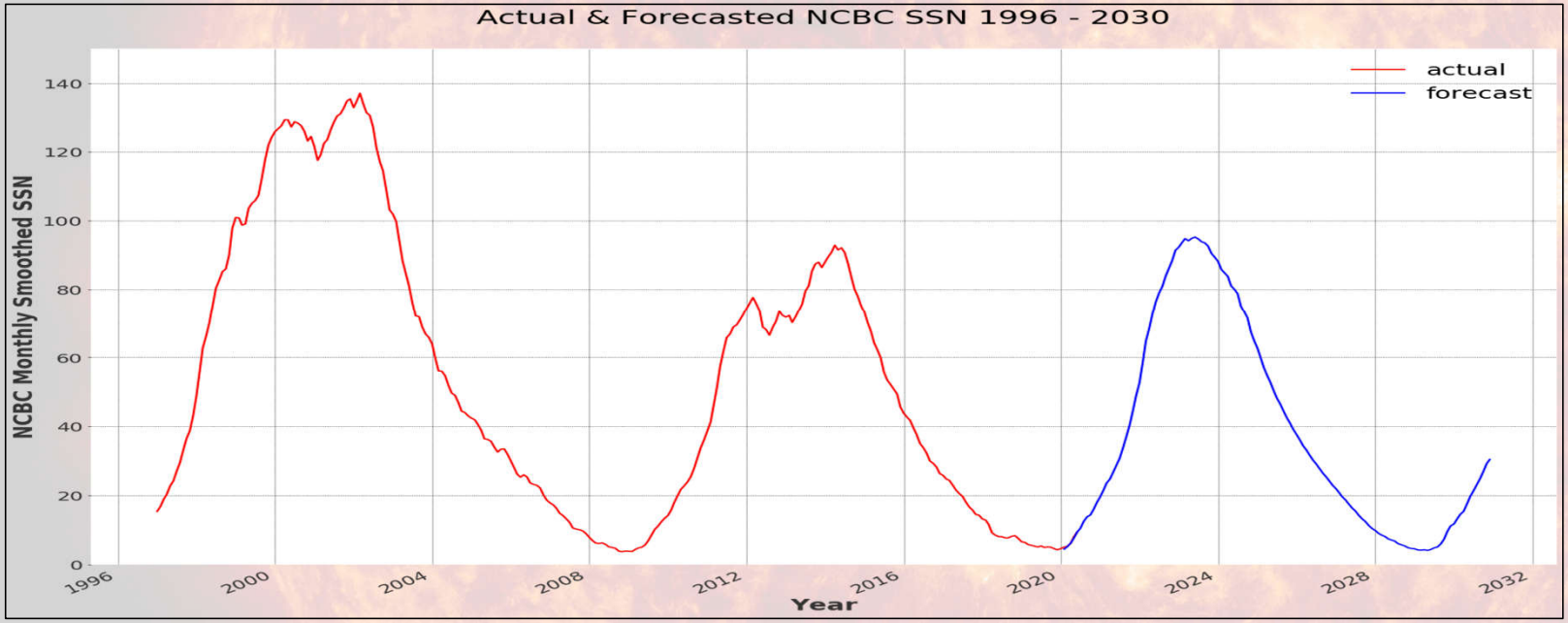
N-BEATS model was imported from darts for training. The basic N-BEATS model parameters used are described below:

- input_chunk_length – The length of the input sequence fed to the model in months.
- output_chunk_length – The length of the forecast of the model in months.
- random_state – Controls the randomness of the weights initialization.
- batch_size – Number of time series (input and output sequences) used in each training pass.
- n_epochs – Number of epochs over which to train the model.

| Model Parameters | Values |
|---|---|
| input_chunk_length | 12 |
| output_chunk_length | 3 |
| n_epochs | 100 |
| random_state | 15 |

## FORECASTED SUNSPOT NUMBER

- In the figure below blue line represents the predicted data whereas the red line shows actual data points. The forecasted graph in blue shows that the maximum SSN amplitude of the 25th solar cycle is 95.243 (±0.32) with a peak in June 2023 (±three months). Compared with previous cycles, the strength of the 25th solar cycle will be similar to solar cycle 24.




- RMSE was computed to be 3.21

# Takeaways

- We have developed a **sunspot detection algorithm and a novel deep neural network based timeseries forecasting model** with the combined goals of detecting sunspots and predicting the strength of the Solar Cycle

- It is demonstrated that the detection algorithm is capable of detecting not only the sunspots present on the solar images but **separately the umbra and penumbra regions** as well. Results of our algorithm were validated with two internationally produced reliable datasets and a very good agreement was found with both of them.

- The algorithm is able to calculate the **area of umbra and penumbra both**.

- It has performed very well compared to the other deep neural network based models with a very low RMSE.

# Takeaways

- The forecasts indicate that Solar Cycle 25 will be similar to slightly stronger than Solar Cycle 24 with a maximum sunspot number of 95.243 (±3.21) and the cycle reaching its peak in June 2023 (± three months). **Our forecast falls within the uncertainty of (Upton & Hathaway, 2018) and the NOAA/NASA forecasts.**

- This is consistent with the consensus forecast of a similar to solar cycle 24 cycle ahead. The proposed method **can be applied to any univariate time series data** that exhibits properties such as trend and seasonality.

- **One limitation is we cannot expect to forecast time series data too far** into the future. When we feed a period of forecast as input back into the model the errors tend to accumulate and get worse over time.


THANK YOU!