EE708: Fundamentals of Data Science and Machine Intelligence

Term Project

EVALUATION

- Each term paper will be evaluated based on the **report (20%), presentation (30%), code (20%), and individual effort (30%).**
  - Report
    - The report is limited to **2 pages**, with an additional page allowed for references.
    - Report should be in IEEE conference template, either **Microsoft Word or LaTeX**. The template can be accessed here: IEEE Template.
    - The report must include the objective, preprocessing steps, model architecture, and performance of the final model.
    - If plagiarism is found, a penalty will be imposed irrespective of the quality of implementation and presentation.
  - Presentation
    - The presentation should be brief and concise.
    - **The number of slides is limited to 5**. It must cover the objective, model architecture, and performance (training and validation).
    - Each group will be given a maximum of **10 minutes** for presentation.
  - Code
    - Use only Python for implementation.
    - You are only allowed to use the training provided data. No external or additional data sources are permitted.
    - Model Design:
      - Do not use pre-trained models or large architecture from existing libraries.
      - You must design your model architecture from scratch using TensorFlow or PyTorch.
      - You may use external libraries for data pre-processing.
      - While you cannot use code others write, you may take inspiration from research papers.
    - Along with the code, you must share a trained model that will be evaluated on a separate test dataset. The test dataset structure will be identical to the training dataset. For this create an evaluation script that must
      - Load and print the model along with its summary.
      - Accept the test dataset path as input.
      - Compute the required performance metrics (check under tasks).
    - Models will be assessed based on **performance metrics and parameter efficiency**.
  - Individual Efforts
    - Questions will be asked during the presentation and code demo to ascertain individual contributions and understanding.
- Final submission consists of the following files:
  - Code.ipynb
  - Model.h5/pth/pt
  - Evaluation script.ipynb (Check the attached example script)
  - Report.pdf

TASK ALLOTMENT

Group no.	Roll Number	Student Name	Allotted task
Group 1	230312	Ch V Sai Koushik	T8: Company Bankruptcy Prediction
	230330	Chilamakuri Kundan Sai	
	230612	Macha Mohana Harika	
	230317	Challa Kethan	
	231033	Srijani Gadupudi	
Group 2	220801	Pranshu Thirani	T7: Book Recommendation System
	220148	Anish Sahu	
	220207	Arpita Chaurasia	
	220387	Edha Bansal	
	220187	Anushka Meena	
Group 3	230636	Marada Teja Satvik	T2: Lung Cancer Detection
	230401	Eswar Naveen Teja Bojja	
	230956	Shashi Bhidodiya	
	230432	Habeeb Ramith Kumar	
	230742	Pasala Bosu Akil Teja	
Group 4	220185	Anurag Gupta	T1: Facial Expression Recognition
	220443	Harshit Sharma	
	220140	Aniket Kumar Choudhary	
	220279	Bali Yaswanth Naidu	
	220252	Atul Kumar Bhongade	
Group 5	220366	Dhruv Varshney	T4: Speech Emotion Recognition
	220405	Gautam Arora	
	220386	Dwij Om Oshoin	
	221102	Sumit Kumar	
	220612	Manas Ranjan	
Group 6	220301	Burri Ganesh Sri Vathshava	T9: Yeast Protein Localization Sites Clustering
	221110	Suryansh Dwivedi	
	230275	Ayushi Mishra	
	230354	Devansh Abhay Dhok	
	230293	Bhavnoor Singh	
Group 7	231018	Soma Koushik	T5: Music Genre Classification
	231174	Voora Rakesh	
	230425	Gudi Praneeth Sai	
	230290	Bharatula Anirudh Srivatsa	
	230916	Sanjay Raghav Vangala	

Group no.	Roll Number	Student Name	Allotted task
Group 8	231066	Suyash Kapoor	T4: Speech Emotion Recognition
	230899	Saksham Verma	
	230187	Archita Goyal	
	230191	Aritra Ambudh Dutta	
	230464	Harshpreet Kaur	
Group 9	241050	Sujal Kumar	T11: Sentiment Analysis
	241230009	Okesh Choudhary	
	241230012	Sanjay Singh Shekhawat	
	241040101	Getiso Gelato Tuloro	
Group 10	210182	Archit Agarwal	T11: Sentiment Analysis
	210096	Akshay Choudhary	
	210227	Astha Tibrewal	
	210241	Avni Maheshwari	
	220736	Norah Sharan Srivastava	
Group 11	220226	Aryan Mittal	T5: Music Genre Classification
	220437	Harshit	
	220522	Keshav Khandelwal	
	220056	Adhiraj Gupta	
	220560	Kumar Gaurav Prakash	
Group 12	230310	Cezan Vispi Damania	T1: Facial Expression Recognition
	230393	Durbasmriti Saha	
	230443	Harsh Agrawalla	
	230941	Saurav Raj	
	231020	Someshwar Singh	
Group 13	211094	Tadiboina Naga Gowtham	T6: Concrete Compressive Strength Prediction
	210077	Ajay Sankar Makkena	
	210662	Nelluru Mourya Reddy	
	230568	Koneti Karthik	
	210937	Sarvasiddi Lakshmi Ruthika Ram	
Group 14	220947	Sambuddha Chakrabarti	T10: SMS Spam Detection
	220678	Nagisetty Vinay	
	221130	Tanmay Soni	
	220763	Pawan Dhakar	
	220495	Kanav Singh Chouhan	
Group 15	210492	Karan Mundhra	T3: Land Use Classification
	210711	Patil Amol Sanjiv	
	210311	Deepanshu	
	210847	Rishav Dev	
	210395	Govinda	

Group no.	Roll Number	Student Name	Allotted task
Group 16	231040414	Suraj Jaiswal	T5: Music Genre Classification
	242040402	Akshay Raina	
	241040034	Kuldeep Chaudhary	
	241040002	Aditya Raj	
	241040087	Suryansh Singh	
Group 17	240118	Anant Aggarwal	T1: Facial Expression Recognition
	240340	Devansh Chaturvedi	
	240265	Bharat Sharma	
Group 18	220813	Pratyush Gupta	T9: Yeast Protein Localization Sites Clustering
	230715	Om Chandrakant Chaudhari	
	200054	Aditya Sharma	
	220238	Ashutosh Rabia	
Group 19	241040099	Piyush Tiwari	T6: Concrete Compressive Strength Prediction
	241040100	Ravi Kumar	
	231040607	Ayushi Ojha	
	241180013	Vaddadi Namrata	
	241040068	Richik Majumder	
Group 20	241040030	Jayesh Shailendra Upadhyay	T3: Land Use Classification
	241040049	Muhammed Anas M	
	241040069	Rishi Chaturvedi	
	241040407	Souvik Atta	
	241040083	Soumyadip Bera	
Group 21	220558	Kuldeepak Dhar Dwivedi	T8: Company Bankruptcy Prediction
	220251	Atharva Singh	
	220709	Nikhil Jain	
	220639	Mayank Jhunjhunwala	
	220369	Dileep Gurjar	
Group 22	241110611	Vijiyant Tanaji Shejwalkar	T2: Lung Cancer Detection
	241040404	Rishabh Bhat	
	242040404	Pankaj Kumar Barman	
	242040604	Amit Kumar Sharma	
	241040405	Rishikesh Chandrashekhar Malkar	
Group 23	220939	Saksham Parihar	T11: Sentiment Analysis
	220354	Dharvi Singhal	
	220954	Samyak Jain	
	220899	Riya Agarwal	
Group 24	241040078	Sayan Datta	T10: SMS Spam Detection
	241040063	Priyanshu Kumar Bhushan	
	241010041	Pankaj Singh	
	230550	Kavadi Rakesh	

Group no.	Roll Number	Student Name	Allotted task
Group 25	210391	Gautam Raghuvanshi	T10: SMS Spam Detection
	210252	B P Hitesh	
	210030	Abhinav Mittal	
	210934	Sarthak Agarwal	
	210350	Divij Singla	
Group 26	210618	Mohd Amir Khan	T4: Speech Emotion Recognition
	210258	Banothu Mithun Raj	
	218070575	Maligireddy Anjali	
	220321	Daksh Dua	
	220694	Nandini Vaid	
Group 27	210358	Divyansh Mittal	T9: Yeast Protein Localization Sites Clustering
	210377	Gandhi Khush Chandreshkumar	
	210705	Parthapratim Chatterjee	
	211044	Snehal Shridhar Kane	
	221116	Swarna Raj	
Group 29	240354	Dhruv Gupta	T8: Company Bankruptcy Prediction
	240870	Rishit Dutta	
	240319	Daksh M Jain	
	241040080	Shashwat Amit Parikh	
	230606	Lokesh Kumar	
Group 30	210881	Rupesh Kumar Meena	T2: Lung Cancer Detection
	210204	Aryan Srivastava	
	210121	Amit Kumar	
	241010053	Sai Vishnu Prasath	
	201032	Swapnil Bagde	
Group 31	230576	Krishna Kumayu	T7: Book Recommendation System
	230677	Naman Agarwal	
	200099	Aman Kumar Meena	
	230759	Poorvie Sadagopan	

TASKS DESCRIPTION

Data Type	Task
Images	**T1: Facial Expression Recognition** **Objective:** Classify grayscale images of human faces into one of seven emotional categories: Angry, Disgust, Fear, Happy, Sad, Surprise, or Neutral. This is relevant for applications in human-computer interaction, mental health monitoring, and automated feedback systems. **Dataset:** The training set consists of 28,709 grayscale images of faces labeled by emotion. **Metric:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
	T2: Lung Cancer Detection **Objective:** Detect and classify lung cancer types based on CT scan images into four categories: adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal (non-cancerous) lung tissue. This supports early cancer detection and treatment planning. **Dataset:** The training set comprises 684 CT scan images labeled by cancer type. **Metric:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
	T3: Land Use Classification **Objective:** Classify land use types into 21 categories based on aerial imagery, supporting research in urban planning, environmental monitoring, and resource management. The dataset includes categories such as agricultural, forest, freeway, river, and tennis court, among others. **Dataset:** The dataset contains a total of 1,680 images, with 80 images for each of the 21 classes. Each image is 256x256 pixels. **Metric:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
Speech/Audio	**T4: Speech Emotion Recognition** **Objective:** Recognize emotions from speech audio files by analyzing vocal characteristics and patterns. This is useful in virtual assistants, emotion-aware systems, and therapeutic applications. **Dataset:** The training set includes 1,140 audio files (19 speakers × 60 files per speaker). Metadata for audio files provides information on modality, vocal channel, emotion, intensity, statement, repetition, and actor attributes. **Metric:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
	T5: Music Genre Classification **Objective:** Classify audio files into one of 10 music genres based on their audio features. This supports personalized music recommendations and content categorization. **Dataset:** The training set consists of 800 audio files, with 80 files per genre across 10 genres. **Metric:** Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
Heterogeneous (numerical, categorical)	**T6: Concrete Compressive Strength Prediction** **Objective:** Predict concrete compressive strength using 8 input features related to mixture components and curing conditions. This is crucial for civil engineering applications, ensuring the safe and optimal use of materials in construction. **Dataset:** The training set includes 824 instances with 8 input features and 1 output variable. **Metric:** Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE).

Data Type	Task
Heterogeneous (numerical, categorical)	**T7: Book Recommendation System** **Objective:** Analyze user preferences and book characteristics to recommend relevant and engaging books. This system aims to enhance user experience. **Dataset:** The training set comprises a dataset of user interactions, including user demographics, book information, and user ratings. **Metric:** Precision at k=10, Recall at k=10, and F1-Score. These metrics evaluate the accuracy of the top-10 recommendations.
	T8: Company Bankruptcy Prediction **Objective:** Predict company bankruptcy using multiple business features, where bankruptcy is defined based on business regulations. This aids in financial risk assessment and economic stability analysis. **Dataset:** The training set contains 5,455 instances with multiple features related to business health. **Metric:** Accuracy, Precision, Recall, F1-Score.
	T9: Yeast Protein Localization Sites Clustering **Objective:** Cluster proteins into groups based on their attributes to identify localization patterns within cells. This task is essential for understanding protein functions and cellular organization. **Dataset:** The yeast dataset contains 1,187 instances with attributes such as sequence recognition scores and other discriminant analysis measures. **Metrics:** Adjusted Rand Index (ARI), Normalized Mutual Information (NMI)
Text	**T10: SMS Spam Detection** **Objective:** Classify SMS messages as either spam (unwanted) or ham (legitimate). This ensures efficient spam filtering and user convenience. **Dataset:** The training set includes 4,457 SMS messages labeled as spam or ham. **Metric:** Accuracy, Precision, Recall, F1-Score.
	T11: Sentiment Analysis **Objective:** Analyze movie reviews to classify their sentiment as either positive or negative. This assists in opinion mining and decision-making for consumer insights and market analysis. **Dataset:** The dataset contains 25,000 movie reviews labeled with binary sentiment polarity. Reviews are stored in directories as text files named based on unique identifiers and ratings. **Metric:** Accuracy, Precision, Recall, F1-Score.