KCDH Proposed Interdisciplinary Dual Degree Projects - 2022-2023

1. Investigations into Mental Health

- **Supervisor:** Pushpak Bhattacharyya

The two project topics are related. We would like to analyze the text, speech and images of people (patients) for an assessment of their emotional state, personality type etc. using NLP, Speech and Computer Vision techniques. The model will be primarily multimodal, multitasking deep neural nets.

Relevant recent publications:

(1) T. Saha, S. Reddy, S. Saha and P. Bhattacharyya, Identifying Mental Health Disorders from Counseling Conversations, IEEE Transactions on Computational Social Systems (IEEE TCSS), Accepted.

(2) Soumitra Ghosh, Swarup Roy, Asif Ekbal and Pushpak Bhattacharyya, CARES: CAuse Retrieval for Emotion in Suicide notes, ECIR, Stavanger, Norway, 10-14 April 2022.

(3) Chanchal Suman, Aditya Gupta, Sriparna Saha and Pushpak Bhattacharyya, A Multimodal Personality Prediction System, Knowledge-Based Systems Journal (KBS), Accepted.

---

2. Investigations into Personality Types

- Supervisor: Pushpak Bhattacharyya

The two project topics are related. We would like to analyse the text, speech and images of people (patients) for an assessment of their emotional state, personality type etc. using NLP, Speech and Computer Vision techniques. The model will be primarily multimodal, multitasking deep neural nets.

Relevant recent publications:

(1) T. Saha, S. Reddy, S. Saha and P. Bhattacharyya, Identifying Mental Health Disorders from Counseling Conversations, IEEE Transactions on Computational Social Systems (IEEE TCSS), Accepted.

(2) Soumitra Ghosh, Swarup Roy, Asif Ekbal and Pushpak Bhattacharyya, CARES: CAuse Retrieval for Emotion in Suicide notes, ECIR, Stavanger, Norway, 10-14 April 2022.

(3) Chanchal Suman, Aditya Gupta, Sriparna Saha and Pushpak Bhattacharyya, A Multimodal Personality Prediction System, Knowledge-Based Systems Journal (KBS), Accepted.

---

3. Topological data analysis based classification of pathological gait in humans

- **Supervisor:** Debasish Chatterjee
- **Co- Supervisor:** Neeta Kanekar

Gait dynamics refers to the magnitude of stride-to-stride changes and the corresponding temporal evolution of the limbs and torso over time. There is reasonable evidence to suggest that the alterations in the locomotor control system of human beings due to neurodegenerative diseases reflect in certain structural changes of their gait. A critical study of the human gait dynamics is, therefore, central to predicting, on the one hand, the onset of nervous disorders that affect locomotion, and on the other hand, evaluating clinical conditions and improving the physiological conditions of neuro-degenerative patients. This project targets the development of a (nonlinear) framework based on topological data analysis for the study of the human gait dynamics.

---

4. A phase III randomized study to develop a novel artificial intelligence based prognostic tool using PET-CT images and pathology images

- **Supervisor:** Dr. Hasmukh Jain
- **Co- Supervisor:** Prof. Ganesh Ramakrishnan

We plan to develop a novel AI-based model using the PET CT and pathology images at baseline that can prognosticate and help choose the treatment protocol.

---

5. Applications of AI/ML in Neurosurgery

- **Supervisor:** Prof. Dipti Gupta

Neurosurgeons usually rely on their experience and clinical evidence to make a decision and provide a prognosis. Regardless of how well-trained or experienced a neurosurgeon is, manual handling of information/data remains challenging for the human capacity. Furthermore, the presence of atypical cases, and lack of access to trained experts adds up to addressing the neuro-issues where time is of crucial importance to provide neurosurgical interventions. There are multiple requirements of AI/ML in pre-, intra- or post-operative neurosurgery as discussed with a neurosurgeon, who is collaborating with our group. In this project, we would be aiming to develop an assistive AI device that helps neurosurgeons in the localization of the tumors or seizure zone.

---

6. NMR-AI based Hybrid Approach for Modelling the Progression of Type 2 Diabetes Mellitus

- **Supervisor:** Prof. Ashutosh Kumar
- **Co- Supervisor:** Prof. Manjesh K Hanwal

Type 2 Diabetes Mellitus (T2DM), a metabolic disorder - has become a primary global health concern. The world's second diabetic capital, India, is expected to witness a ~71.7% rise in diabetic cases by 2045. Currently used glucose level determination methods can report the disease long after the damage has been done and cannot be reversed. Therefore, in this proposal, we aim to decipher the metabolic patterns of individuals using high-throughput techniques like NMR and Mass Spectrometry. This data will then be subjected to different machine learning algorithms to generate a metabolic fingerprint for the early signature of diabetes. Then, fingerprints will be utilized to predict potential markers in individuals much before the onset of the disease. This personalized diagnosis will also be providing founding stones for personalized treatment. Diabetes is a metabolic disease, and metabolism being highly unique for every individual. Thus, personalized diagnosis, prevention, and diabetes treatment are needed.

---

7. Radiation reduction algorithms in computed tomography

- **Supervisor:** Prof. Prof. Ajit Rajwade

The amount of radiation administered to patients, especially neonatal, geriatric or pregnant patients, and inadvertently to doctors and technicians, is a common concern in computed tomography (CT) scanning. The total radiation can be lowered by reduction in the number of tomographic projections acquired by a CT scanner (termed the "few views setting") and/or reduction in the XRay beam power per projection (termed the "low dosage setting". However this results in poorer quality image reconstruction, which could seriously interfere with accurate diagnosis. In many procedures, multiple CT scans per patient are acquired, either across weeks as in cancer surveillance, or in quick succession during CT-guided surgical procedures such as CT-guided biopsy or ablation. In these applications, the changes to the patient's anatomy across scans is restricted to relatively smaller areas. If the initial few scans are acquired with high radiation dosage, they can therefore act as guiding templates to enable more accurate reconstruction of the subsequent scans acquired from significantly low dosages. However while doing so, one must accurately determine the regions of genuine anatomical change across the scans, directly from the projection measurements. The aim of this project is to use machine learning algorithms to enable superior image reconstruction using the readily available prior information in such repeated scanning procedures. The long term goal is to encourage CT scanner manufacturers to incorporate machine learning algorithms at the backend to enable more intelligent image scanning at lower dosages without sacrificing image quality. For more details see the following: Preeti Gopal, "Tomographic Reconstruction: A Radiation Reduction Approach", PhD Thesis, IIT Bombay, January 2020, link:

https://drive.google.com/file/d/1QsrKGRW9RVx9z_sPUJR-e3Uwzqvvw9q9/view?usp=sharing

Yadnyesh Patil, "Tomographic reconstruction with low radiation dosage

or view undersampling with prior templates", MTech Thesis, IIT Bombay, July 2021, link: https://drive.google.com/file/d/1EopTPN6BtuYyFGmvbEQG6Upy7hdgBPql/view?usp=sharing

---

8. Designing predictive model for Childhood Congenital Glaucoma using Clinical and Genetic data

- **Supervisor:** Prof. Abir De
- **Co-Supervisor:** Prof. Ganesh Ramakrishnan

Despite being rare, childhood glaucoma is a disease of increasing concern in india. In many cases, they are detected after significant manifestation of the disease. In this work, we wish to devise predictive models which will predict the possibility of childhood glaucoma based on his/her clinical characteristics. More specifically, we aim to predict two types of glaucoma --- congenital glaucoma and acquired glaucoma, where the latter is often driven by eye surgery of young children. To do so, we would first invest in collecting the genetic signals from patients. Then, we will develop predictive models which will use the genetic signals and other clinical records to predict the possibility of development of childhood glaucoma in a patient. Finally, we will use our model to recommend the type of surgery the clinician should adopt to treat the patient, if needed.

---

9. Internship Project mentioned against ICICI Lombard on Predicting Health Insurance cost, Predicting disease/hospitalization using fitness and health data

- **Supervisor:** Prof. Usha Ananthakumar

---

10. Constructing a multi-dimensional healthcare cost index

- **Supervisor:** Prof. Siuli Mukhopadhyay
- **Co-Supervisor:** Prof. Souvik Banerjee

The aim of this project is to create a healthcare cost index based on individual-level and community-level healthcare expenditures and also factors like demographic characteristics and socio-economic and health status, that may affect the cost of healthcare in India. The developed index will enable insurance companies and actuarial scientists to assess healthcare expenditure in different domains, compare costs across different population strata based on age, economic status, rural and urban living, and study the benefits of the numerous healthcare schemes and programs available in India. The index will be built using recent data, including pandemic time, using health behaviors along with population parameters, health state, and healthcare costs, thus allowing a study of pandemic effects on healthcare expenditures. We will use both econometric/statistical techniques as well as AI/ML methods to build the healthcare cost index.

---

11. Automated feature selection for biomarker discovery from big biological data

- **Supervisor:** Prof. Pramod Wangikar
- **Co-Supervisor:** Prof. Siuli Mukhopadhyay

Proteins and metabolites, the new class of biomarkers are expected to bring a paradigm shift in the diagnosis, monitoring and treatment of human disease and will make personalized medicine a reality in near future. Moreover, the next generation biomarkers are likely to be based on the inference drawn from multiple metabolite or protein molecules rather than single measurements such as the blood glucose level that is currently used for the diagnosis of diabetes. The present project focuses on the discovery of biomarkers from big biological data involving genomics, proteomics or metabolomics. Our primary objective will be to select the best combination of biomarkers or minimum subset of features to predict class labels such as disease vs. healthy or one category of disease vs. another. The challenges include: (i) small amount of labelled data, (ii) unbalanced data with too many features and too few samples, (iii) the need to learn to predict classes that may have only subtle differences, and (iv) high degree of inherent biological variability (unrelated to the class label) and instrument noise. We will use a number of feature selection methods and machine learning tools together with the concepts of multitask learning to achieve the task of biomarker discovery. A large number of public domain databases are available that will be used as test cases. See a recently published paper to understand the broad objectives of the proposed project (https://www.frontiersin.org/articles/10.3389/fgene.2019.00452/full). You may also see literature on multitask machine learning. The co-supervisor is an expert in data science while the supervisor is active in the fields of metabolomics, systems biology and big biological data analysis. The candidate must have a strong foundation in AI/ML/data science apart from the willingness to work in cross-disciplinary areas. The supervisors will make active efforts to arrange a summer internship for the student for the summer between the 4th and 5th years, subject to approval by the faculty advisors. This will be either at a company or with one of our collaborators.

---

12. Applications of machine learning in analyzing higher-order mass spectrometry metabolomics data

- **Supervisor:** Prof. Pramod Wangikar
- **Co-Supervisor:** Prof. Amit Sethi

Liquid chromatography coupled with mass spectrometry (LC-MS) is a very popular technique that can be used to simultaneously analyze thousands of proteins or metabolites in biological samples. The latest LC-MS instruments acquire higher dimension data that includes orthogonal properties such as chromatographic retention time (RT), mass to charge ratio (m/z), tandem MS data (MS2) and collision cross sectional area (CCS values). With such high dimensionality, data analysis becomes quite cumbersome. Further, the signals need to be ascribed to known compounds based on matches with the compound libraries. Further, experimentally determined values of MS2 and CCS are available only for a

limited number of compounds compared to the available chemical space. Thus, with the limited labelled data, data-efficient machine learning tools need to be developed including quantitative structure-property relationships (QSPR) in the domain of metabolomics. The following articles may be explored that describe similar problems in the domain of interest. https://www.nature.com/articles/s41467-020-18171-8 or https://www.nature.com/articles/s41467-021-21352-8 The co-supervisor is an expert in machine learning while the supervisor is active in the fields of metabolomics, systems biology and big biological data analysis. The candidate must have a strong foundation in AI/ML apart from the willingness to work in cross-disciplinary areas. The supervisors will make active efforts to arrange a summer internship for the student for the summer between the 4th and 5th years, subject to approval by the faculty advisors. This will be either at a company or with one of our collaborators.

---

13. Proteomics and big data analysis of breast cancer

- **Supervisor:** Sanjeeva Srivastava

---

14. Smart digital platform to assess and manage child's physical, cognitive and immune health through nutritional intervention

- **Supervisor:** K V Venkatesh

A digital platform is envisaged wherein the physiological details are captured in order to assess child's physical, tilized to develop such a platform. A scientific research backed nutritional suggestion will be incorporated to provide intervention to improve overall child's wellness acognitive and immune status. A state of the art methods from the fields of systems biology and AIML will be und growth.

---

15. Autonomous Air Ambulance / Casualty Evacuation system

- **Supervisor:** Dhwanil Shukla
- **Co-Supervisor:** Rajkumar Pant

An autonomous drone-based casualty evacuation system is being envisaged. The DDP shall focus on first identifying the type of care that must be provided to the patient in transit and then designing/coming up with ways to provide the care inside the drone without additional human involvement.

---

16. Classification of chest pain using ECG & EMG signals.

- **Supervisor:** Dr. Nirmal Punjabi
- **Co-Supervisor:** Prof. Ganesh Ramakrishnan

Chest pain are observed in our day to day lives and it is a matter of concern due to its conventional association with heart problems. But, it is also associated with many non-cardiac reasons like occupational disorders due to bad posture, or other non-cardiac reasons. The aim of this project is to analyze ECG & EMG data to give preliminary differentiation between different types of chest pain.

---

17. A framework for high-fidelity digitization of historical medical records

- **Supervisor:** Prof Ganesh Ramakrishnan

The objective of the proposed project is to achieve the following:

Create a collection of digital, searchable copies of historical medical records. The method to do this include features:

1. Automatic character recognition based text from scanned images

2. Post-editing for OCRed texts

3. Searchable PDF documents with image and text layers

Development and application of the necessary technologies that are required to build the following:

1. AI-based OCR technologies to scan the historical medical records as images and turn them into searchable text

2. An interactive platform for post-edit OCR texts to correct errors that occurred during the OCR process

---

18. Optical Character Recognition(OCR) and Name-Entity Recognition for integration with Bahmni EMR for Patient Medical Document

- **Supervisor:** Prof Ganesh Ramakrishnan

---

19. A novel Artificial Intelligence-based prognostic approach using PET-CT images and pathology images for advanced-stage Hodgkin Lymphoma

- **Supervisor:** Prof Ganesh Ramakrishnan

For advanced Hodgkin's lymphoma, two chemotherapy regimens are often used for treatment, (i) ABVD and (ii) BEACOPP regimens. ABVD regimen has comparatively low cure rates as compared to the BEACOPP regimen, but it is tolerated well in most patients. BEACOPP, though associated with better outcomes, comes with the risk of significant toxicities. In this study, we propose to use AI/ML to evolve a strategy to help select the protocol based on the patient's disease characteristics at baseline using PET-CT images and pathology images. This can help avoid the comparatively toxic BEACOPP protocol where ABVD can give equally good results. The study includes 1) a retrospective aspect, where we use the previously available datasets to develop AI models to predict the outcome of ABVD treatment, 2) the prospective aspect, where we will use the AI models to predict the response to the treatment of ABVD & BEACOPP chemotherapy protocols.

---

20. OCR and Applications in Indian Languages

- **Supervisor:** Prof Ganesh Ramakrishnan
- **Co-Supervisor:** Prof. Parag Chaudhuri

The objective of this project is to develop technology to enable the development of applications and open up opportunities that use Indian language OCRs for Electronic Medical Record Systems. OCR for Indian languages is quite challenging due to the richness in inflexions such as differences in scripts, various combinations of conjunct characters for similar sounding words, etc., Even after a good accuracy in OCR, the detected text will need a lot of improvement. Further, in the digitization process of such texts, the second step would be spelling and error correction. Hence, the end goal is to convert the generated OCR text in accordance with the scanned images for printed documents. We will use State-of-the-Art (SOTA) Deep Neural Network (DNN) object detection models such as Faster-RCNN, RetiaNet to train and detect the layout. And With human-in-the-loop, we will have tools and methods for automatically and for users to manually correct the erroneous OCR characters.

---

21. A study to develop an artificial intelligence algorithm for predicting the outcomes of febrile neutropenia (FN) in adolescents and adults with haematological malignancies.

- **Supervisor:** Prof Amit Sethi
- **Co-Supervisor: Prof. Hasmukh Jain**

Febrile Neutropenia (FN) is one of the most frequent and serious complications of chemotherapy. A wide spectrum of pathogens is known to cause FN. A specific microbial agent is, however, isolated only in 23% of the cases1. Delay in initiating appropriate antibacterial therapy results in high rates of mortality and morbidity. IDSA and ECIL guidelines developed in the west help in the selection of antibiotics at the beginning and during the course of FN. These guidelines are largely empirical which leads to the use of an excess of antibiotics. Artificial intelligence can be employed to predict the risk of complications and select the choice of antimicrobials. It can be done by utilizing the high number of data available from medical records and optimizing the big data to train with deep learning methods. We propose to conduct a study to develop an artificial intelligence algorithm to predict the risk of FN and select antibiotics in cancer patients receiving chemotherapy.

---

22. Improving disease diagnosis with human-machine collaboration

- **Supervisor:** Prof. Abir De

In a wide range of high stake and critical applications, societies rely on the judgement of human experts to make consequential decisions, which have significant impacts. However, due to the number of decisions to be taken as well as the shortage of experts, the timeliness and quality of such decisions are often compromised in practice. For example, in high-demand medical applications, patients in several countries need to wait for months to get an appointment by a physician; in content moderation, online publishers often temporarily prevent their customers from commenting in their portals, etc.

In recent years, machine learning models have surpassed human performance in many applications e.g., machine translation, image recognition, etc. However, in critical applications where a small error can make consequential impacts as exemplified above, humans are preferred over ML models, as the latter can incur significant predictive errors. However, in most cases, the ML models are not aware of the presence of humans and therefore, they are trained for full automation. In this work, we will see how we can build machine learning models under triage--- a new machine learning setup--- where machines and humans together achieve better performance than what they can achieve at an individual level. We design methods on how tasks can be distributed across machines and humans so that they perform at their comfort zone and do better. Moreover, we show that introducing an auxiliary machine improves predictive performance of the model. Experiments on a wide variety of supervised learning tasks using synthetic and real data show that our approach outperforms those provided by several competitive baselines.

---

23. Designing predictive model for Childhood Congenital Glaucoma using Clinical and Genetic data

- **Supervisor:** Prof. Abir De

Childhood glaucomas account for a major form of blindness in children in developing nations, especially Primary Congenital Glaucoma (PCG, 1 in 10,000 live births in the West, 1 in 3300 live births in Andhra Pradesh as per Eye Diseases Study). We propose to devise predictive models using clinical and genetic data (n = 100) which will (a) primarily aim to identify genetic signals that can correlate with the clinical features (which can enable genetic counselling for expectant parents among surviving children in their 20s) (b) predict the surgical success and/or visual prognosis in children with PCG, and (c) exploratively predict the type of surgery that may be beneficial in PCG associated with specific genetic signals. This pilot project will lead to algorithms that can help clinicians counsel parents of children with blinding glaucoma for an overall better outcome.

---

24. Integrate protein-DNA interaction network with genomic data to correlate and predict the impact of mutations in clinical strains of M. tuberculosis.

- **Supervisor:** Prof. Sarika Mehra

In spite of significant medical advances, tuberculosis (TB) is a major killer worldwide. In India alone, more than 12 lakh people are diagnosed with tuberculosis every year, with 2.7 lakh resulting in death. 40% of these cases are children. The major problem with conventional tuberculosis therapy has been the emergence of multi-drug resistance (MDR) and extensively drug resistance (XRD) Mycobacterium tuberculosis strains. Whole genome sequencing of many clinical strains show a number of mutations present in these clinical strains. In this project, we will represent the regulatory network of this bacteria in a Boolean framework. To determine the Boolean function at each node, gene expression data will be utilized along with in-house developed algorithms. The model will be simulated to predict the impact of various mutations that are seen in clinical strains of drug-resistant M. tuberculosis.

---

25. Automatic landmark detection in medical imaging

- **Supervisor:** Prof. Darshan Shah

---

26. Molecular Dynamics Simulation of various proteins involved in rare endocrine diseases

- **Supervisor:** Prof. Ambarish Kunwar

---

27. Investigating the association between cellular morphometric traits and chemoresistance in cancer cells

- **Supervisor:** Prof. Rajdip Bandyopadhyaya

Tumor is formed of multiple sub-population of cells indicating a highly heterogeneous mass. Intra-tumoral heterogeneity greatly complicates the study of molecular mechanisms driving cancer progression and thereby our ability to predict patient outcomes. Heterogeneity might arise due to cell to cell variation in genome, transcriptome or proteome. Furthermore, recent findings have shown that cellular heterogeneity is associated with cancer progression, metastasis and chemotherapy resistance. Thus, it is necessary to understand the heterogeneity of the tumor to predict disease outcome and thereby design cancer treatment. One of the prominent phenotype based on which the cellular heterogeneity can be recognized is cell morphology (in terms of cellular surface area, perimeter, circularity, aspect ratio etc. based on images of cancer cells).

In this work, we propose to explore cell morphology to predict the sensitivity of the pancreatic cancer cell to chemotherapy. For this existing images and their measurements will be analyzed by algorithms and learning tools being used by the joint proposing team (Chemical and Electrical Engg.) of this project. Further, we will determine the heterogeneity with the cancer cells and co-relate with chemoresistance. For these we will investigate the following:

(i) Different traits of cell morphology will be measured pre- and post- chemotherapy. Then the cells will be segregated into different groups based on unsupervised hierarchical clustering of the cell morphology traits. This will help in understanding how the subpopulation within tumor varies in terms of cellular morphology. Further, degree of morphological heterogeneity as measured by Shannon's entropy of the morpho-types will be explored.

(ii) Chemoresistant cells will be developed using drug as selection pressure, based on our already existing collaboration with Advanced Centre for Treatment, Research and Education in Cancer (ACTREC), Navi Mumbai. These can be used for the proposed work, where these developed cells will be clustered into different subpopulations based on cell morphology, in turn co-relating morphometric measurements with chemoresistance.

(iii) Co-relation measurement will be carried out to determine which particular types of subpopulation are more predominant in imparting chemoresistance to tumor. This will also help us in understanding the evolution of cell morphology during the acquirement of chemoresistance.

28. Using machine learning methods to predict important residues determining substrate specificity of efflux pumps in antimicrobial resistance

- **Supervisor:** Prof. Sarika Mehra

  Efflux pumps play an important role in imparting anti-microbial resistance to pathogens. The experimental identification of substrates of transporters is costly and time-consuming. Thus, the development of robust bioinformatics-based methods for the prediction of membrane transport proteins and their substrate specificities is an important task. In this project, we will use various properties of the protein sequence of efflux pumps to build a classification model to predict important residues that may impart substrate specificity.

---

29. Cognitive and Emotional markers of Autism

- **Supervisor:** Prof. Rashmi Gupta

  The project aimed to examine the cognitive and emotional markers in individuals/children with Autism.

---

30. Using Multiresolution Machine Learning in the analysis of biomedical data, images and series of biomedical images

- **Supervisor:** Prof. Vikram M. Gadre

  Multiresolution techniques, multi-rate digital signal and image processing, wavelets and time-frequency methods have been used traditionally in the analysis of biomedical signal and image data. The growth of techniques in machine learning and neural networks has inspired the view that a union of these paradigms is the 'way to go', in the future in dealing with biomedical signal and image data. The purpose of this IDDDP Project is to further that endeavour, already being pursued in my R&D group.