# IHub-Data, IIIT Hyderabad

**TIH on Data Driven Technologies (Data Banks, Services & Analytics)**

# Success Story: Dataset collection efforts and Technology development (COVID, and Cancer) for developing the early diagnostic warning/solutions and mass screening.

As part of the Technology Innovation Hub (TIH) initiative, we have launched few activities in various dimensions such as medical diagnostics, disease risk stratification, decision support systems, drug designand discovery, and proactive disease prevention and health maintenance. IHub is focused on putting together large-scale datasets to help develop solutions through applied research. Therefore, our strategy is to create and curate large datasets, develop algorithms, and bring the solutions to the field through technology start-ups. We have distributed our activities along four major focus areas, namely, Cancer diagnostics, Neuro and Mental Health, Data Driven Drug Discovery, and Public Health.

In Cancer Diagnostics, we plan to collect datasets of the most common cancers in India such as oral, breast, lung, GI, cervical cancers to develop AI driven screening tools for early detection, tumour subtypes / grades / stages diagnosis, survival prediction (longitudinal data), and response to treatment (longitudinal data). One such effort is ‘Artificial Intelligence (AI) driven predictive models for assessing cancer probability in Oral, Breast, and Lung regions.’ This is one of our first initiatives in the healthcare domain that is centred around image-based early screening and diagnosis of cancer wherein our research group worked constantly towards the development of the app that facilitated and simplified the GRACE Foundation’s (NGO in the Cancer space) data collection effort. A total of 501 patients' data was collected in Phase one. Dr. Chinna Babu (Oncological Surgeon) and his team from GRACE Foundation have now initiated the first level of sanity check with help of an annotation portal that is developed at IHub. Prof. Vinod, who is leading the cancer diagnostics focus group and his team is developing an AI model that uses the data collected to predict the probability of cancer in a low resource setting and to develop an early screening tool for oral cancer. Cancers such as oral have favourable outcomes, if they are detected early and treated. Hence the current effort is placing huge emphasis on early screening as well as developing a mobile solution for screening camps in rural settings. The developed tool which makes a diagnosis based on a photograph of the oral cavity will facilitate deployment of the solution with the help of semi-skilled staff, thereby mitigating the requirement of a trained clinician to be available in the rural camps. Further, the aim is to take the solution to the field and not wait for patients to show up in the secondary and tertiary centres when it might be too late!

Globally there has been an interest and increased adoption of advances in Computer Vision in diagnostics. Prof. Vinod P.K. of IIIT-H DATA I-HUB Foundation, in collaboration with NIMS Hyderabad is working on curating an Indian Pathology Dataset (IPD) where a large-scale data collection of histopathology images from Indian patients along with their clinical observations are collected into a repository. Samples are being procured from several other pathology departments reputed hospitals in Telangana and Andhra Pradesh. IPD will be used to develop AI solutions for accelerated and accurate diagnosis of several types of cancer. This machine learning solution can help by assisting a pathologist in making quality diagnosis as well as in quickening the time- consuming manual process of scoring a whole slide image. Initially, the team is focused on digitizing the existing glass slides to create an Indian dataset of various malignancies using a digital whole slide scanner. This project enables distant reporting, telepathology by merely sharing a link of the histopathology or cytopathology image to the doctor for diagnosis. This will not only make the diagnostics more accessible in the remote areas or low resource settings but also be easier to get a second opinion too!

# 

Another significant paradigm shift will be witnessed in the archival and retrieval process of the histopathological slides. Medical institutions archive the slides for about 5-15 years. This requires a lot of space and infrastructure. The quality of staining also deteriorates with time and sometimes the glass slides break too! With digital archiving, the images can be barcoded and easily accessible whenever required. Installing expensive scanning machines in every hospital may not be feasible, the iHub-Data team is in preliminary discussions to have a centralised scanning facility housed at IIITH, serving as a one-stop shop for sample collection and digitization. The repository will be useful in future for pathology training too.

# Covid-19 has posed severe Public Health challenges. And with the virus affecting different people differently, it has become imperative to predict the progression of the disease. Our research team in collaboration with Intel Corp and CSIR-IGIB have collected data from Gandhi Hospital, Hyderabad and Max Group of Hospitals, New Delhi and developed machine learning models to categorise risk and predict mortality in Indian patients. The team has successfully carried out the research and published their work in Frontiers in Public Health, “Machine Learning Based Clinical Decision Support System for early Covid-19 Mortality Prediction” wherein the team has attempted to provide a mortality prediction as early as 16 days before the outcome. Our team has conducted the study on 544 Covid positive patients admitted in the MAX group of hospitals in New Delhi. Using the same dataset of Covid positive patients from Wuhan, the team has identified 5 biomarkers that can be used to predict mortality with 96% accuracy. Two different machine learning methods were developed for risk stratification and mortality prediction, based on the features extracted from the Indian Cohort and both of which have yielded exceptionally good results. Some of these features have been identified as predictors of the progression of the covid-19 disease. This early prediction model will help healthcare professionals to accelerate the decision-making process for appropriate treatments. Our team has also analysed the viral genome collected from the Indian patients to find mutations that may be associated with severity of the Covid-19 disease and carried out comparative studies to understand the difference

between Indian and Wuhan cohort. This study is believed to help accelerate the decision-making process in healthcare system for focused and eﬃcient treatment.

Graphical user interface, website

Description automatically generated