**Mental Health Survey Project**

**Deep Learning Model to Predict Depression from Survey Data**

**1. Introduction**

Depression represents a growing public health challenge worldwide. Early detection plays a crucial role in enabling timely intervention and effective treatment strategies. This comprehensive project aims to develop a predictive model that can accurately identify individuals who may be at risk of experiencing depression based on various factors including:

* **Demographic information**
* **Academic background**
* **Lifestyle patterns**
* **Medical history**

Our solution employs advanced deep learning techniques to identify hidden patterns within survey data and delivers predictions through a user-friendly Streamlit web application hosted on AWS infrastructure.

**2. Problem Statement**

**Primary Objective**

Build a robust predictive model that classifies individuals into **high-risk** or **low-risk** categories for depression.

**Key Requirements**

* **Fairness**: Ensure equitable performance across all demographic groups (gender, age, profession, city)
* **Accuracy**: Achieve high predictive accuracy using balanced datasets
* **Practicality**: Create a solution suitable for real-world applications in:
  + Healthcare systems
  + Corporate wellness programs
  + Government policy-making
  + Educational institutions

**3. Dataset Overview**

**Data Source**

* **Format**: Mental Health Survey Data (CSV)
* **Target Variable**: Depression (Binary: 0 = No, 1 = Yes)

**Feature Categories**

**Demographics**

* Age
* Gender
* City

**Academic & Professional**

* Educational Degree
* Current Profession
* Work/Study Status

**Stress Indicators**

* Academic Pressure
* Work-related Stress
* Financial Pressure

**Lifestyle Factors**

* Dietary Habits
* Sleep Duration
* Work/Study Hours per Day

**Medical & Family History**

* Family History of Mental Illness
* History of Suicidal Thoughts

**4. Data Preprocessing Pipeline**

**4.1 Data Cleaning & Standardization**

* **Categorical Field Corrections**: Fixed invalid entries in City, Profession, Degree, Sleep Duration, and Dietary Habits
* **Column Optimization**: Removed irrelevant identifiers (Name, ID)

**4.2 Missing Value Treatment**

* **Numerical Features**: Median imputation
* **Categorical Features**: Mode imputation

**4.3 Feature Engineering & Transformation**

* **Categorical Encoding**: One-hot encoding for all categorical variables
* **Data Transformation**: Log transformation applied to skewed features:
  + Academic Pressure
  + CGPA
  + Study Satisfaction

**4.4 Class Balancing**

* **Technique**: SMOTE (Synthetic Minority Oversampling Technique)
* **Strategy**: Stratified by Age Group × Profession to maintain demographic representation

**4.5 Feature Scaling**

* **Method**: StandardScaler for all numerical features
* **Purpose**: Ensure optimal neural network performance

**5. Model Development**

**5.1 Technical Framework**

* **Primary Framework**: TensorFlow with Keras API
* **Model Type**: Deep Neural Network for Binary Classification

**5.2 Architecture Design**

**Input Layer**

* **Features**: 100+ encoded features

**Hidden Layers Architecture**

| **Layer** | **Neurons** | **Activation** | **Regularization** |
| --- | --- | --- | --- |
| Layer 1 | 256 | ReLU | Batch Normalization + Dropout (0.15) |
| Layer 2 | 128 | ReLU | Batch Normalization + Dropout (0.20) |
| Layer 3 | 64 | ReLU | Batch Normalization + Dropout (0.25) |
| Layer 4 | 32 | ReLU | Batch Normalization + Dropout (0.30) |

**Output Layer**

* **Neurons**: 1
* **Activation**: Sigmoid (Binary Classification)

**5.3 Training Configuration**

* **Loss Function**: Binary Crossentropy
* **Optimizer**: Adam (Learning Rate: 0.001)
* **Evaluation Metrics**: Accuracy, Precision, Recall, AUC
* **Early Stopping**: Patience = 10 epochs

**6. Model Evaluation Results**

**6.1 Overall Performance Metrics**

**Validation Results (23,999 samples)**

* **Accuracy**: 97%
* **Precision**: 0.97 (Class 0), 0.97 (Class 1)
* **Recall**: 0.97 (Class 0), 0.96 (Class 1)
* **F1-Score**: 0.97 (Class 0), 0.96 (Class 1)
* **AUC**: 0.9946

**Test Results (48,728 samples)**

* **Accuracy**: 96%
* **Precision**: 0.96 (Class 0), 0.96 (Class 1)
* **Recall**: 0.96 (Class 0), 0.96 (Class 1)
* **F1-Score**: 0.96 (Class 0), 0.96 (Class 1)
* **AUC**: 0.9940

**6.2 Demographic Fairness Analysis**

**Gender-Based Performance**

| **Gender** | **Sample Size** | **Accuracy** | **Precision** | **Recall** |
| --- | --- | --- | --- | --- |
| Female | 24,708 | 97% | 0.96-0.97 | 0.96-0.97 |
| Male | 28,913 | 96% | 0.96-0.97 | 0.96-0.97 |

**Age Group Performance**

| **Age Group** | **Sample Size** | **Accuracy** | **F1-Score** |
| --- | --- | --- | --- |
| <18 | 839 | 86% | 0.77-0.90 |
| 18-29 | 8,325 | 86% | 0.86-0.87 |
| 30-44 | 17,170 | 97% | 0.97 |
| 45-59 | 22,394 | 100% | 1.00 |

**6.3 Professional Category Analysis**

**Student Population (7,047 samples)**

* **Accuracy**: 87%
* **Precision**: 0.90 (Class 0), 0.84 (Class 1)
* **F1-Score**: 0.86-0.87

**Working Professionals (41,681 samples)**

* **Accuracy**: 98%
* **Precision**: 0.97 (Class 0), 0.99 (Class 1)
* **F1-Score**: 0.98

**6.4 Profession-Specific Performance**

**Exceptional Performance (>98% Accuracy)**

* **Digital Marketer**: 100% accuracy
* **Teacher** (11,914 samples): 99% accuracy, F1-score 0.99
* **Pharmacist**: 99% accuracy
* **HR Manager**: 99% accuracy
* **Chef, Electrician, Manager**: 99% accuracy each

**Strong Performance (95-98% Accuracy)**

* **Software Engineer**: 97% accuracy
* **Content Writer**: 99% accuracy
* **Student**: 87% accuracy (largest challenging group)

**6.5 Educational Background Analysis**

| **Degree Type** | **Accuracy** | **Performance Level** |
| --- | --- | --- |
| PhD, ME, MED | 98-99% | Exceptional |
| MBBS, MD | 97% | Strong |
| BE, BTECH | 97% | Strong |
| BBA, MBA | 97% | Strong |
| CLASS 12 (7,143 samples) | 95% | Good |

**6.6 Geographic Distribution Analysis**

**Metropolitan Cities**

* **Major Metros** (Mumbai, Delhi, Bangalore, Chennai): 97-98% accuracy

**Tier-2 Cities**

* **Secondary Cities** (Pune, Ahmedabad, Hyderabad): 96-98% accuracy

**Smaller Cities**

* **Regional Centers**: 96-98% accuracy maintained

**7. Deployment Architecture**

**7.1 Streamlit Application Features**

**User Interface Components**

* **Mental Health Assessment Page**: Interactive form for real-time risk predictions
* **Business Use Cases Dashboard**: Specialized modules for different stakeholders

**Stakeholder Modules**

* **Healthcare Providers**: Clinical decision support
* **Mental Health Clinics**: Patient prioritization tools
* **Corporate Wellness Programs**: Employee monitoring dashboard
* **Government/NGOs**: Population analytics interface

**7.2 Cloud Infrastructure**

* **Platform**: Amazon Web Services (AWS)
* **Services**: EC2 instances
* **Benefits**:
  + **Scalability**: Auto-scaling based on demand
  + **Reliability**: High availability architecture
  + **Security**: Enterprise-grade data protection

**8. Results & Strategic Insights**

**8.1 Model Performance Highlights**

* **Exceptional Overall Accuracy**: 96-97% across all test scenarios
* **Demographic Equity**: Consistent performance across gender and professional groups
* **Age-Related Patterns**: Higher accuracy observed in middle-aged groups (30-59 years)
* **Student Population Insight**: Lower performance indicates need for targeted interventions

**8.2 Clinical and Practical Applications**

**Healthcare Providers**

* **Early Screening**: 96% accuracy for preliminary assessments
* **Resource Optimization**: Efficient patient triage

**Mental Health Clinics**

* **Queue Prioritization**: Demographic-aware risk stratification
* **Treatment Planning**: Data-driven intervention strategies

**Corporate Wellness Programs**

* **Employee Monitoring**: Proactive mental health support
* **Risk Assessment**: Workplace stress identification

**Educational Institutions**

* **Student Support**: Targeted intervention programs
* **Campus Wellness**: Population-level monitoring

**Government & NGOs**

* **Policy Planning**: Population-level risk analytics
* **Resource Allocation**: Evidence-based decision making

**8.3 Model Reliability Indicators**

* **Cross-Validation Consistency**: 97% validation vs 96% test accuracy
* **Balanced Performance**: Equal precision and recall across classes
* **Demographic Robustness**: Stable performance across 30+ professions and 25+ cities

**9. Project Deliverables**

**9.1 Technical Assets**

* **Source Code Repository**: Complete preprocessing, training, and deployment scripts
* **Trained Model Files**:
  + mental\_health\_survey\_final.keras
  + Feature scaler objects
  + Feature column specifications

**9.2 Application Components**

* **Streamlit Web Application**: Interactive user interface with demographic-aware predictions
* **AWS Deployment Package**: Production-ready hosting configuration

**9.3 Documentation & Evaluation**

* **Comprehensive Performance Analysis**: Detailed metrics across all demographic segments
* **Deployment Guide**: Step-by-step implementation instructions
* **User Manual**: Application usage guidelines

**10. Future Enhancement Roadmap**

**10.1 Advanced Analytics**

* **Longitudinal Analysis**: Track model accuracy and performance over extended periods
* **Temporal Pattern Recognition**: Identify seasonal or cyclical depression risk factors

**10.2 Customization & Localization**

* **Regional Adaptation**: City-specific model fine-tuning
* **Cultural Considerations**: Local demographic factor integration

**10.3 Professional Specialization**

* **Occupation-Based Risk Profiles**: Industry-specific intervention strategies
* **Workplace Integration**: Corporate-specific deployment modules

**10.4 Technology Expansion**

* **Mobile Application**: Smartphone-based accessibility and real-time monitoring
* **API Development**: Integration capabilities for existing healthcare systems
* **Real-Time Analytics**: Streaming data processing for continuous assessment

**Conclusion**

This Mental Health Survey Project represents a significant advancement in predictive mental health analytics, combining cutting-edge deep learning techniques with practical deployment strategies. The model's exceptional performance (96-97% accuracy) across diverse demographic groups demonstrates its potential for real-world clinical and administrative applications.

The comprehensive evaluation across multiple dimensions—demographic, professional, educational, and geographic—ensures that the solution is both effective and equitable, making it suitable for widespread implementation in healthcare, corporate, and governmental contexts.