CHAPTER 2

Theory and Literature review

2.1 Introduction

Steels are used in the construction and fabrication of engineering structures, with service temperatures ranging from subzero to about 600°C over long periods of time. The vast majority of iron alloys are ferritic because they are cheap and it is easy to modify their microstructures to obtain an impressive range of desirable properties.

The fabrication of steels unavoidably involves welding, a complex process incorporating numerous metallurgical phenomena. It is not surprising therefore, that the final microstructure both inside the weld metal and in all adjacent regions affected by welding heat, is remarkably varied. Many of the important features of weld microstructure can now be calculated using a combination of thermodynamics and kinetic theory [4]. Such calculations are now being performed routinely in industry during the course of alloy design or when investigating customer quaries.

Naturally, it is the mechanical properties of the weld which enter the final design procedures. There has been some progress in estimating the yield strength from the microstructure using combinations of solution strengthening, grain size effects, precipitation hardening and dislocation strengthening [4]. The ultimate tensile strength can in a limited number of cases be calculated empirically from the yield strength [1]. However, there has been no progress at all in creating models for vital properties such as ductility, toughness, creep and fatigue strength [3].

2.2 Ferritic Steels

2.2.1 Heat Resistant Steels

Steels are used widely in the construction of power plant. They have to resist creep deformation, oxidation and corrosion. The superheater pipes carrying steam from boilers to high pressure(HP) turbines typically experience steam at 565°C under 15.8 MPa pressure and are made of low-alloy steels. In HP turbines the rotor is fabricated as a single forging of 1Cr-MoV steel. Tempering at 700°C leads to the formation of stable carbides which are distributed uniformly in the ferrite matrix. These carbides improve the creep resistance at the service temperature [2]. Turbine blades experience both erosion and high tensile forces. High strength and corrosion resistant 12CrMoV steel is used in fabrication of turbine blades [5]. The 3½Ni-Cr-Mo-V alloy has good hardenability combined with high strength of about 1100 MPa and good toughness. These steels are air cooled from 870°C and tempered at 650°C. Due to their strength and toughness these materials are used to fabricate the low pressure turbine rotor, which is nearer to the generator. The generator rotor is also fabricated with this material [6].

Table 2.1 .Chemical composition of some steels have been used Power Plant [7], all units are in wt%.

Steel	C	Si	Mn	Mo	Cr	V
2¼Cr-1Mo	0.15	0.50	0.45	1.0	2.25	---
12Cr-1Mo	0.15	0.40	0.6	1.0	12	---
3½Ni-Cr-Mo-V	0.15	0.30	0.70	0.10	1.5	0.11

Cr-Mo Steels

These materials are resistant to corrosion by sulphur products and hence were used first in the petroleum industry. Once their oxidation resistance and high temperature strength were appreciated, they began to be applied in the steam power generating industry. More recently, these steels have been used in fabricating thick pressure vessels. The oxidation resistance and high temperature strength depends on the amount of chromium and molybdenum present in that alloy. Excellent high-temperature (565°C) strength is obtained in 2¼Cr-1Mo steels (Table.2.1), which are generally used in the bainitic condition. A tempering heat-treatment gives the required alloy carbides; the most important are M2C, M7C3 and M23C6, where M represents a metallic element, where M represents a metallic element.

2.2.2 Structural steels

Steels for structural applications are used at ambient temperatures and the main property requirements are strength, ductility and toughness. The vast majority of these steels have a yield strength in the range 300-550 MPa with a mixed microstructure of ferrite and pearlite. These are used in critical applications, such as bridges, buildings or ship construction and may undergo sophisticated themomechanical processing to refine the microstructure and greatly improve the toughness. Such alloys may contain quantities of fine bainite or even martensite when the overall concentration is small.

All structural steels have to be welded. For this reason and to minimize the cost, the total alloy concentration is generally less than 5wt%. The weld metals used for joining structural steels also range in yield strength between 350 and 550 MPa, but can be much stronger (900MPa) for special steels used in the construction of submarines. The preferred weld microstructures contain large quantities of acicular ferrite which, because of its scale and chaotic arrangement, gives good toughness. However, quantities of allotriomorphic ferrite, Widmanstetten ferrite, martensite and retained austenite may also be present.

2.3 Mechanical properties of weld deposits

Many engineering components are fabricated using welding. The integrity of the fabrication is usually asserted on the basis of mechanical properties. Strength, ductility and toughness are considered as the essential mechanical properties. Previous work on the modelling of weld metal mechanical properties is reviewed in this chapter.

2.3.1 Strength

The capacity of a material to withstand static load can be determined using a tensile test, in which a standard specimen is subjected to a continually increasing uniaxial load until it fractures, Fig 2.1. The load-elongation curve is plotted and the results are usually restated in terms of stress and strain, which within reasonable limits are independent of the geometry of specimen, Fig.2.2:

$$\text{Engineering stress,} \quad \sigma E = \frac{P}{A} \qquad (2.1)$$

Engineering strain, $\epsilon E = \frac{Lf - L}{L}$ (2.2)

Where P is load, Ao is initial cross-sectional area and Lo and Lf are initial and final lengths of the sample.

The material at first extends elastically; if the load is released the sample returns to its original length. After exceeding the elastic limit the deformation is said to be plastic, so the sample does not regain its original length if the load is released. With continued loading the engineering stress reaches a maximum beyond which the sample develops a neck. This local decrease in cross-sectional area focuses deformation until the sample fractures.

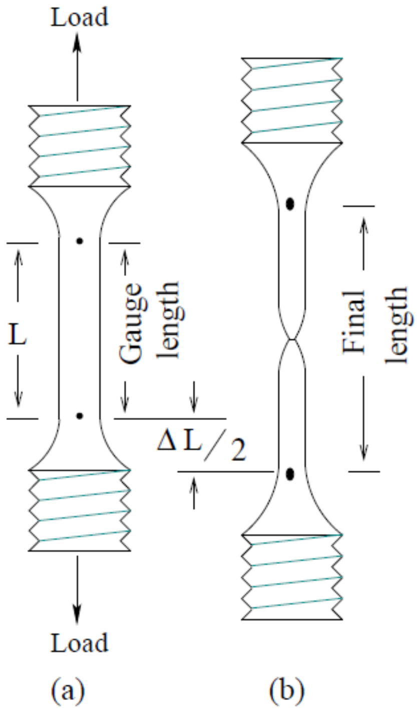


Figure 2.1: schematic diagram of tensile test specimen, a) before testing b) after testing. ΔL is the total extension of the specimen during the tensile test.

The yield stress is defined as the stress at which plastic deformation just starts as the stress-strain curve deviates from linearity. Because of the difficulty in precisely measuring this deviation, a‟‟0.2% proof stress‟‟ is used which is the stress at 0.002 plastic strain, Fig.2.2b.The proof stress is sometimes referred as the „yield stress‟‟. The maximum engineering stress is called the „ultimate tensile stress‟, whereas the stress at which the sample fractures is called the „fracture stress‟.

Engineering stresses and strains do not account for the change in the load bearing cross-sectional area of the sample during deformation. The true stress and strain do so and are defined as follow:

$$\sigma = \sigma E\ (\epsilon E+1) \tag{2.3}$$

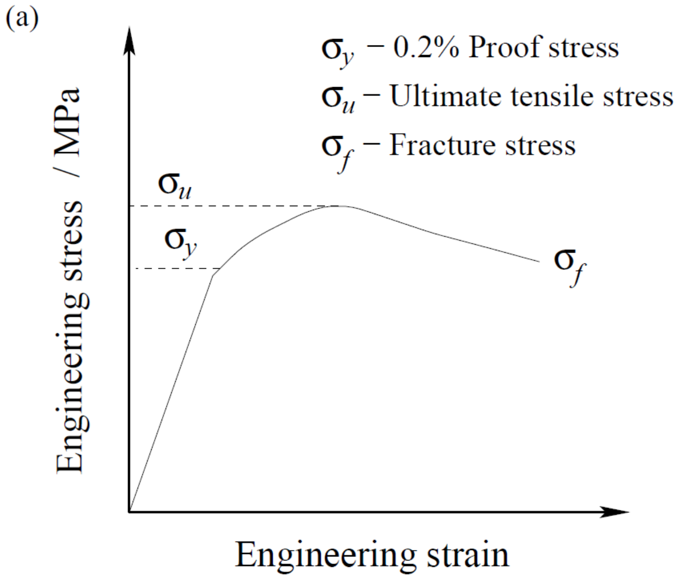
$$\epsilon = \ln\ (\epsilon E+1) \tag{2.4}$$

This leads to a change in the form of the stress-strain plot as illustrated in Fig.2.3.

The engineering strain and true stress are comparable at small strains. The flow curve of many metals as expressed in terms of the true stress σ and true strain ϵ can be represented as:

$$\sigma= K\epsilon \tag{2.5}$$

where „K‟ is value of the flow stress at ϵ=1.0 and „n‟ is the strain hardening exponent. Both these



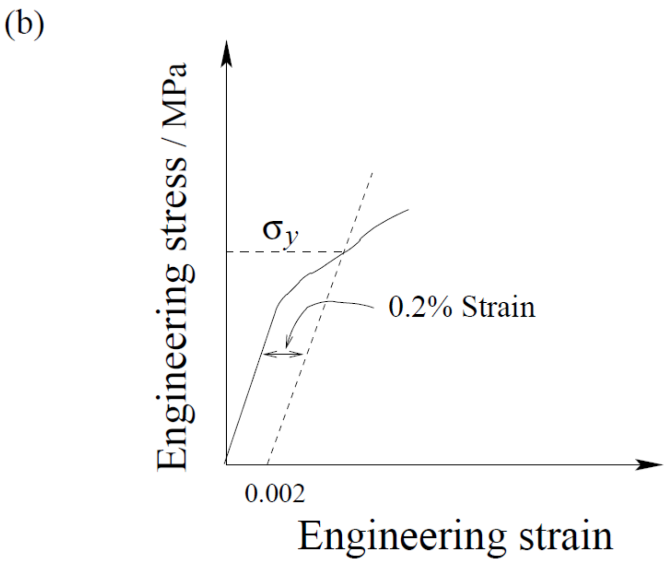


Figure 2.2: engineering stress=strain curve showing a) different stresses, b) 0.2% proof stress.

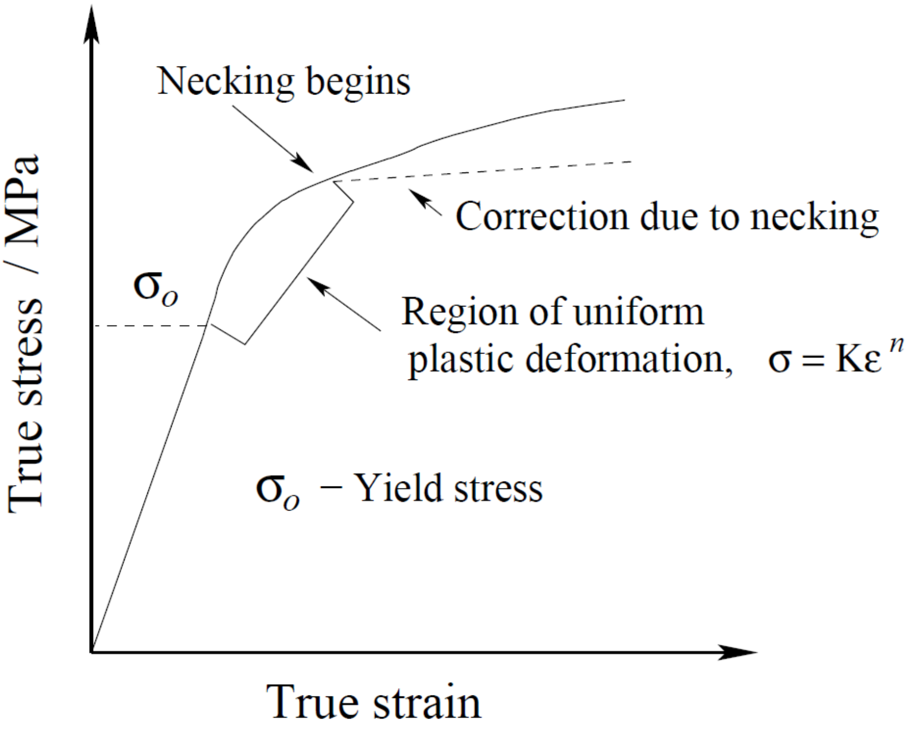

Necking begins
Correction due to necking
σ_o
Region of uniform
plastic deformation, $\sigma = K\varepsilon^n$
σ_o − Yield stress
True stress / MPa
True strain

Figure 2.3: True stress - true strain curve (flow curve).

Parameters can be estimated from a logarithmic plot of true stress and true strain. In practice, the strain hardening exponent may vary with strain but equation 2.5 is nevertheless a useful representation of plastic deformation.

2.3.2.Ductility

Ductility is important because an engineering component should show considerable plasticity before fracture. Ductility, as measured in a tensile test, is usually expressed as elongation or reduction in area:

$$\text{Elongation} = \frac{Lf - L}{L} \qquad (2.6)$$

$$\text{Reduction in area} = \frac{Af - A}{A} \qquad (2.7)$$

Where, Lf is the length of sample at fracture, L is initial length, Ao is the initial area of cross-section and Af is the area of cross-section at fracture. Both elongation and reduction in area are frequently expressed as percentages.

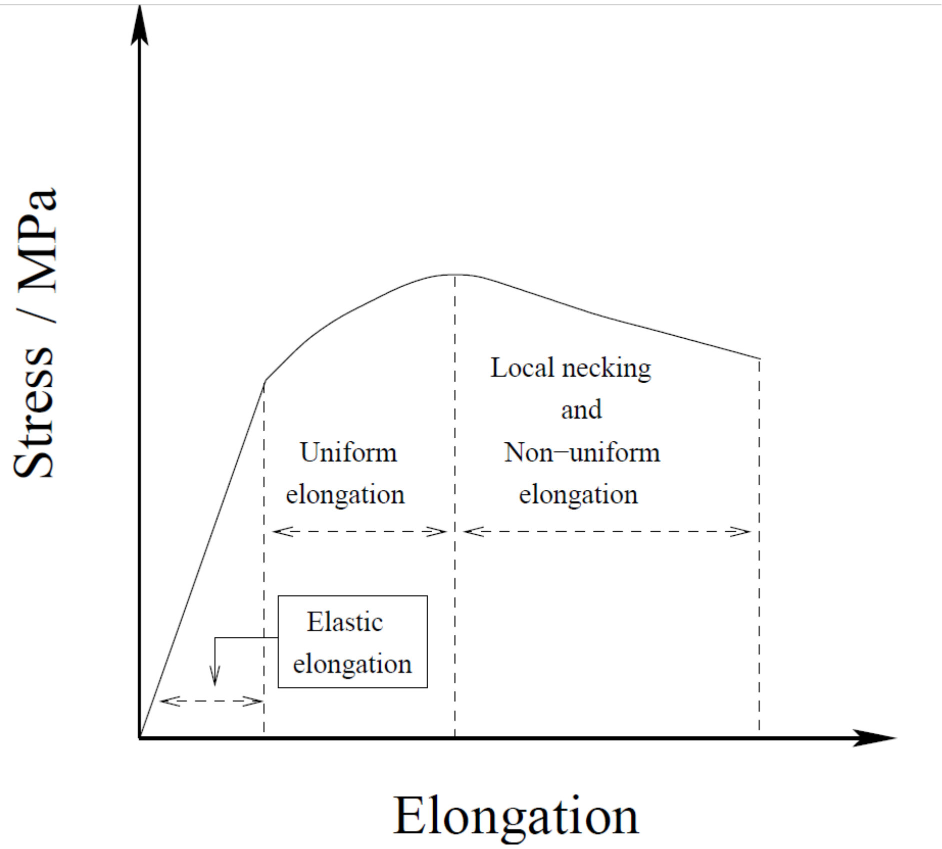


Figure2.4: The stress-elongation curve. The elastic elongation is exaggerated for clarity.

Plastic strain can be subdivided into two components, an initial uniform strain where the cross-section of the sample is identical along the entire gauge length, and a non-uniform component beginning with the onset of necking. Assuming that equation 2.5 is a true representation of deformation, the stress at the point where necking begins is given by $\sigma=K$ [8].

2.3.3 Charpy impact toughness

Toughness is the ability of the material to absorb energy during the process of fracture. The ability to withstand occasional stresses above yield stress without fracturing is particularly desirable in engineering components. The welded joints should resist brittle fracture; therefore, the weld metal should be tough, with a great deal of energy being absorbed during the process of fracture. One of the popular test methods to characterize toughness is the Charpy impact toughness test is which a square sectioned, notched sample(Fig.2.5) is fractured under specified conditions[9].The absorbed energy during fracture is taken as a measure of toughness. However, Charpy impact test values are empirical since these results cannot be used directly engineering design and can be used only to rank samples in research and development experiments. The test is usually conducted on a material over a range of temperatures to reveal any ductile-brittle transition, (Fig.2.5).

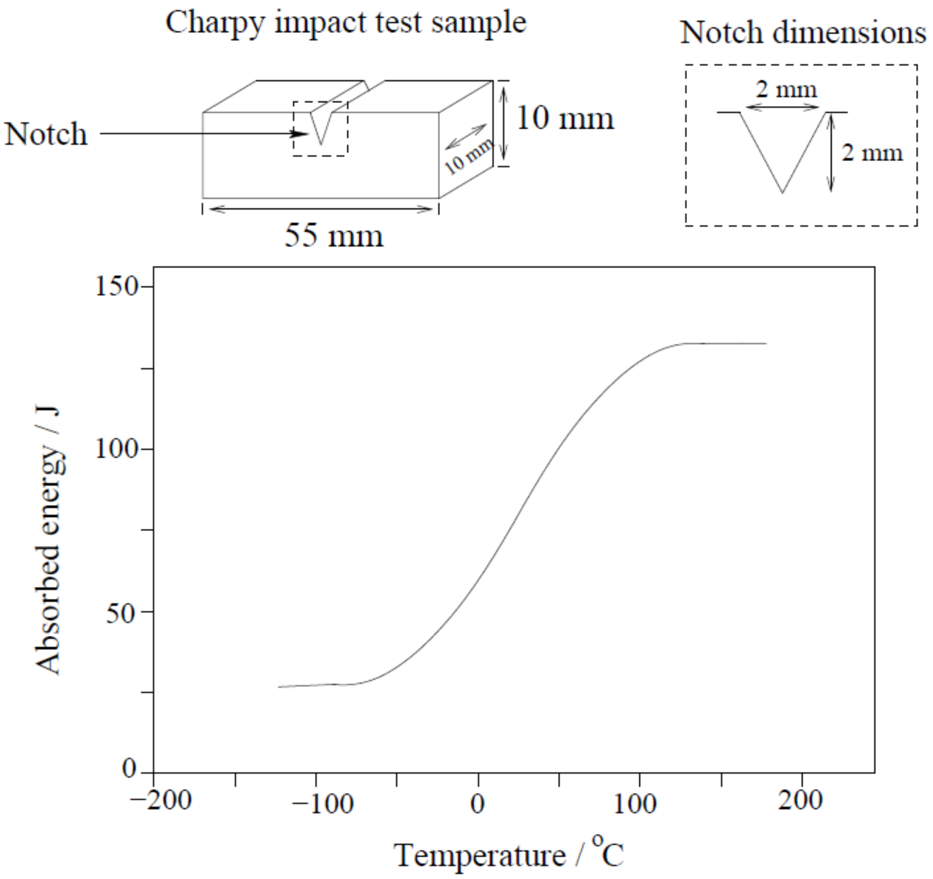


Figure2.5: the Charpy impact test sample and impact toughness versus test temperature curve.

2.4 Arc Welding Processes

Welding is one of the most popular joining methods for steels. The joining of two alloys can be done by melting the two surfaces to be joined by heat, with or without the help of a filler wire. The method by which heat is generated in order to fuse the base metal and filler wire defines the nature of the welding process: electric arc welding, electron beam welding, friction welding. The work presented in this thesis focuses on arc welds which are now described in some more detail.

2.4-1Arc Welding

An electric arc is the source of heat to melt and join metals. As shown in Fig 2.6, an electric arc is struck between the work piece and the electrode which is manually or mechanically moved along the joint or electrode remains stationary while the work piece can be moved. The electrode may or may not be consumed during the process. The molten weld pool is protected by an inert or active gas shroud generated using flux or via an external supply of gases.

2.4.2 Manual Metal Arc Welding

This is also called the shielded metal arc welding (SMAW) process. Its simplicity and versatility makes it popular. A consumable electrode coated with flux (silicates, minerals and metals) is used as shown in Fig 2.7. The coating provides elements which act as arc stabilizers, generate gases and a slag cover to protect the weld pool from the environment and add alloying elements to the weld deposit. The electrode and workpiece are connected to a power source; usually the electrode is connected to the positive terminal of the power source. The arc is initiated by touching the electrode tip to the base metal and then forming an air gap. The heat generated as a consequence melts the base metal, the electrode core and its covering.

2.4.3 Gas Tungsten Arc (TIG) Welding

A non-consumable tungsten electrode is used together with an inert shroud. The key advantage of this process over manual metal arc welding is that higher quality welds can be produced. The equipment used in this process is portable and usable with all metals, for a wide range of thicknesses and in all welding positions.

2.4.4 Gas Metal Arc Welding

Gas metal arc welding (GMAW) uses a continuous wire which is consumed to form the weld metal together with an inert gas shield. The mode of liquid metal transfer from electrode to the base metal can be varied by choosing different types of gases. All metals can be welded by using argon or carbon dioxide. This process gives high weld metal deposition rates and can be automated.

2.4.5 Submerged Arc Welding

As the name indicates, the electric arc and molten weld metal are submerged under a layer of molten flux and unfused granular flux. The tip of a continuously fed consumable wire is the electrode. Because the arc is submerged under molten flux the radiation losses are minimized giving maximum energy efficiency. This is an automated process which can be used with the base metal in the horizontal position.

2.5 Variables Associated with Welding

The most important variables are the process, chemical composition of the weld deposit, heat input, the initial temperature of the base metal at the region to be welded (pre- heating), temperature of the weld deposit during multirun welding (intepass temperature) and heat treatment given to the weld metal after welding (post-weld heat treatment). The type of joint (Fig 2.8) and the material thickness have to be considered in selecting a weld process. The primary function of the heat source is to generate heat to melt the base metal and consumable electrode. The rate of melting is controlled by amount of heat input, denned as:

$$\text{Net heat input}(\text{J mm}^{-1}) = \frac{fIV}{S}$$

where I is the electric current in amperes, V is the voltage applied between power source terminal and electrode expressed in volts, S is the travel speed of the heat source in mm s- 1 and f is the arc transfer efficiency. In most of the arc welding processes the efficiency is between 0.8 and 0.99. The weld metal composition plays a vital role in determining the mechanical properties of the weld joint and the microstructure of weld metal. A post-weld heat treatment is often given to the as- deposited weld to lower the hardness and restore the toughness.

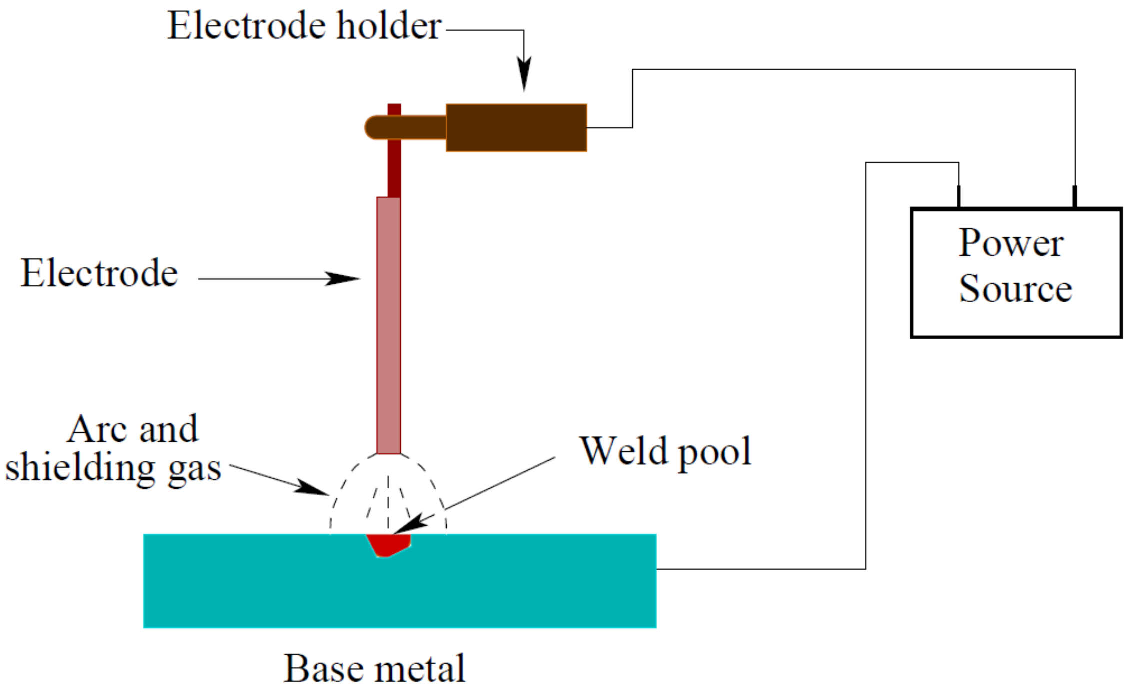


Figure 2.6 Schematic view of arc welding process.

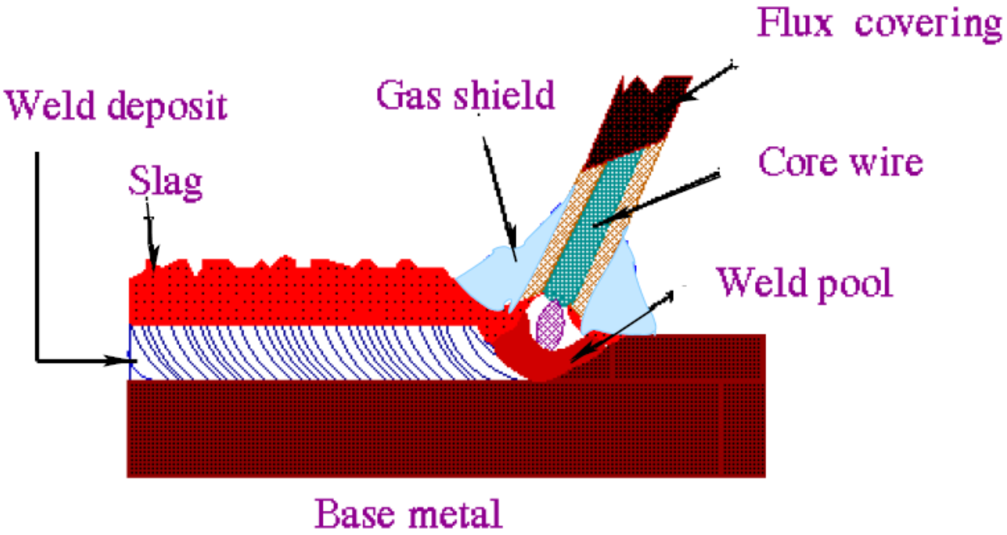


Figure 2.7 Schematic view of manual metal arc welding (MMAW).

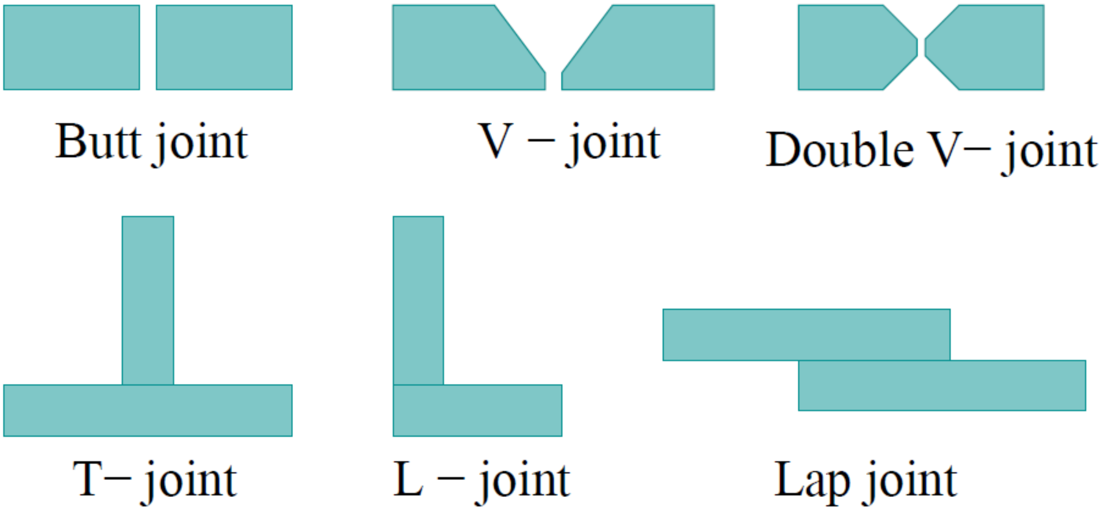


Figure 2.8 Different types of joint preparations.

2.6 Weld Microstructure

When a molten metal solidifies in the gap between components to be joined, this welds the components together. The basic metallurgy of the welded joint can be divided into two major regions: the fusion zone and heat affected zone (HAZ). The fusion zone experiences temperatures above the melting point of the material and represents both the deposited metal and the parts of the base metal melted during welding. The heat affected zone, (Fig. 2.9) on the other hand, represents the close proximity to the weld, where the temperatures experienced are below the melting point and there is a change in the microstructure of the base metal.

2.6.1 Weld Metal Solidification

In steels weld metal solidification starts at edge of the fusion zone into the weld metal with d-ferrite as the initial phase (Fe-Cementite Phase diagram). As it cools, d-ferrite transforms into austenite and with further lowering of the weld metal temperature austenite decomposes to ferrite. Most steels contain small quantities of alloying elements and hence show similar crystal structure changes as pure iron. Therefore in weld metal solidification, weld deposits begin solidification with the epitaxial growth of columnar d- ferrite from the hot grains of the base metal at fusion surface. The grains grow rapidly in the direction of highest temperature gradient and hence show an anisotropic morphology. Those grains with <100> directions parallel to the heat How direction dominate the final microstructure. On further cooling, austenite nucleates and grows along prior d-ferrite grain boundaries, thus adopting the columnar shape of the d-ferrite grains. Fine austenite grains providemore grain boundary nucleating sites; on the other hand coarse grains increase the harden abilityof the weld metal. The columnar shape of the austenite results in few grain boundary junctions when compared with an equi-axed structure. This also contributes to an increase in harden ability.

The cooling rates in the weld metal depend on the distance from the heat source, heat input, inter passtemperature and the geometry of the joint. Because the cooling rates are in practice quite high, weld solidification is a non-equilibrium phenomenon and thus solidification-induced segregation promotes an inhomogeneous microstructure in the weld metal. The amplitude of these concentration and microstructure variations become larger as the alloy concentration increases.

Another important feature, in Flux based welding processes, is non-metallic inclusions. During welding, the flux reacts with atmospheric oxygen and cleans and protects the weld metal by forming oxides and rejecting them into slag. However, the process is not ideal due to convection and rapid solidification, so oxide particles are entrapped in the fusion zone during solidification. These are called slag inclusions, which can serve as nucleation sites within the weld pool. A small volume fraction of inclusions is desirable in welding, as they serve as heterogeneous nucleation sites for councilorferrite. Large fractions are detrimental to the mechanical properties of weld metal.

2.6.2 As-deposited Weld Microstructure

The as-deposited microstructure is that which forms when the liquid weld pool cools to room temperature. This structure contains allotrimorphic ferrite, Widmanst5tten ferrite and councilor ferrite, Fig. 2.10. In a few cases, microstructures containing marten site, banite and traces of pearlite can be found. High- carbon martensite is a hard microstruct ure with low toughness and ductility.

Allotriomorphic Ferrite

Allotriomorphic ferrite (a) usually forms between 1000 and 650"C during cooling of steel weld deposits. Nucleation occurs heterogeneously at the columnar austenite grain boundaries. As the austenite grain boundaries are easy diffusion paths, austenite grain boundaries are decorated with thin layers of anthropomorphic ferrite and the thickness of which is controlled by the diffusion rate of carbon in austenite. In weld deposits, anthropomorphic ferrite appears to grow without the redistribution of substitutional alloying elements during transformation [8]. This mechanism of growth is termed para equilibrium, and occurs as a consequence of the fast cooling rates experienced by welds. In welds, anthropomorphic ferrite is detrimental to the toughness because the continuous network along grain boundaries offers less resistance to crack propagation than councilorferrite [11].

Widmanstàtten Ferrite

This microstructure results from further cooling below the temperature at which anthropomorphic ferrite forms. Primary Widmanst5tten ferrite nucleates directly from the regions of austenite grain boundaries not covered by anthropomorphic ferrite. Secondary Widmanst5tten ferrite nucleates at austenite/ferrite boundaries and grows as sets of parallel plates separated by thin regions of austenite. The austenite remains as retained austenite, or transforms to marten site or pearlite. These latter transformation products are collectively known as micro phases in weld metal terminology, because they are generally present in small fractions. Widmanstatten ferrite is not desirable in weld metals.

Acicular Ferrite

Oxides and non-metallic inclusions serve as nucleation sites for acicular ferrite. Acicular ferrite forms within the columnar austenite grains in competition with Widmanst5tten ferrite. It appears as a fine grained interlocking array of non- parallel laths. The microstructure is highly desirable in welds. The large number of non- parallel grains improves the weld metal toughness by increasing the resistance to crack propagation [12].

Microphases

These are last constituents to form in weld metal. Microphases correspond to the small carbon-rich regions in the weld metal where the last remaining volumes of austenite transform, and consist of mixtures of marten site, carbides, degenerated pearlite, bainite and retained austenite.

2.6.3 Secondary Microstructure

In many circumstances it is difficult to fill the gap at the joint by a single weld pass. Therefore thick sections are welded using many layers of deposited metal, Fig. 2.11. The deposition of each successive layer heat treats the underlying microstructure formed during cooling of the previous run. Some regions of the underlying layers are reheated above the austenitisation temperature, whereas others become tempered. All of the reheated regions contribute to the secondary microstructure.

The Heat Affected Zone

The heat affected zone is the portion of the metal which has not experienced melting, but whose microstructure is altered due to welding heat. There are well-defined microstructures in the heat affected zone as illustrated in Fig. 2.12 . The region immediately adjacent to the fusion boundary is heated to very high temperatures (just below melting temperature) and forms coarse austenite. The austenite grain size decreases sharply with distance from the fusion line and the Rue grained zone will have superior mechanical properties than the coarse grained zone. Moving further away, the peak temperature decreases and will result in partial austenite formation and tempered ferrite in that region; this is called the “partially austerities zone”. The region adjacent to this zone, which is not transformed to austenite, will be tempered.

2.6.4 The microstructure and CCT diagram of weld metals

Those features of weld solidification that are most likely to influence the final microstructure of the weld metal after cooling to ambient temperatures have been presented above. The overall picture is complicated by a number of interacting factors which include:

1. The welding process itself which determines the weld pool size and geometry.
2. The final composition of the melt as influenced by the filler wire, the base metal, fluxes, gases, moisture in the air, etc., and its effect on constitutional supercooling and segregation.
3. The speed of welding and its effect on solidification speeds, crystal morphology and segregation.
4. The weld thermal cycle and its influence on microstructural coarseness and type of transformation product produced during cooling.
5. The effect of weld metal composition, particularly from dilution in high energy welding of microalloyed steels, on precipitation reactions, and especially during reheating or in multi-run welds.

It appears to be unrealistic to attempt to develop CCT diagrams specific to weld metal composition and thermal history. On the other hand, it is useful and informative to express the influence of the various features of welding, as listed above, in a schematic CCT diagram which shows, e.g., the tendency for the C-curves to move to longer or shorter times or the introduction of shape or size changes of the transformation fields, and this is illustrated in Figure 2.13.

Arrows that point left in the diagram denote movement of C-curves to shorter transformation times, and arrows to the right indicate the opposite effect. Thus, austenite stabilizers (e.g. C, N, Mn, Ni, Cu), tend to inhibit transformation, pulling the C-curves towards longer times to transformation. Strong carbide or nitride forming elements (e.g. Mo, Cr, Nb, V, Ti, Al), however, tend to suppress blocky and proeutectoid ferrite, but not acicular ferrite or bainite. Indeed, Nb in particular tends to enhance bainite formation. Slag inclusions, particularly if present in sufficient number and size, also tend to promote the nucleation of acicular ferrite. [10]

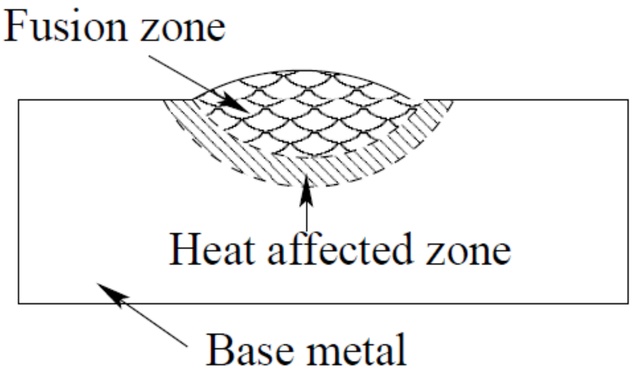
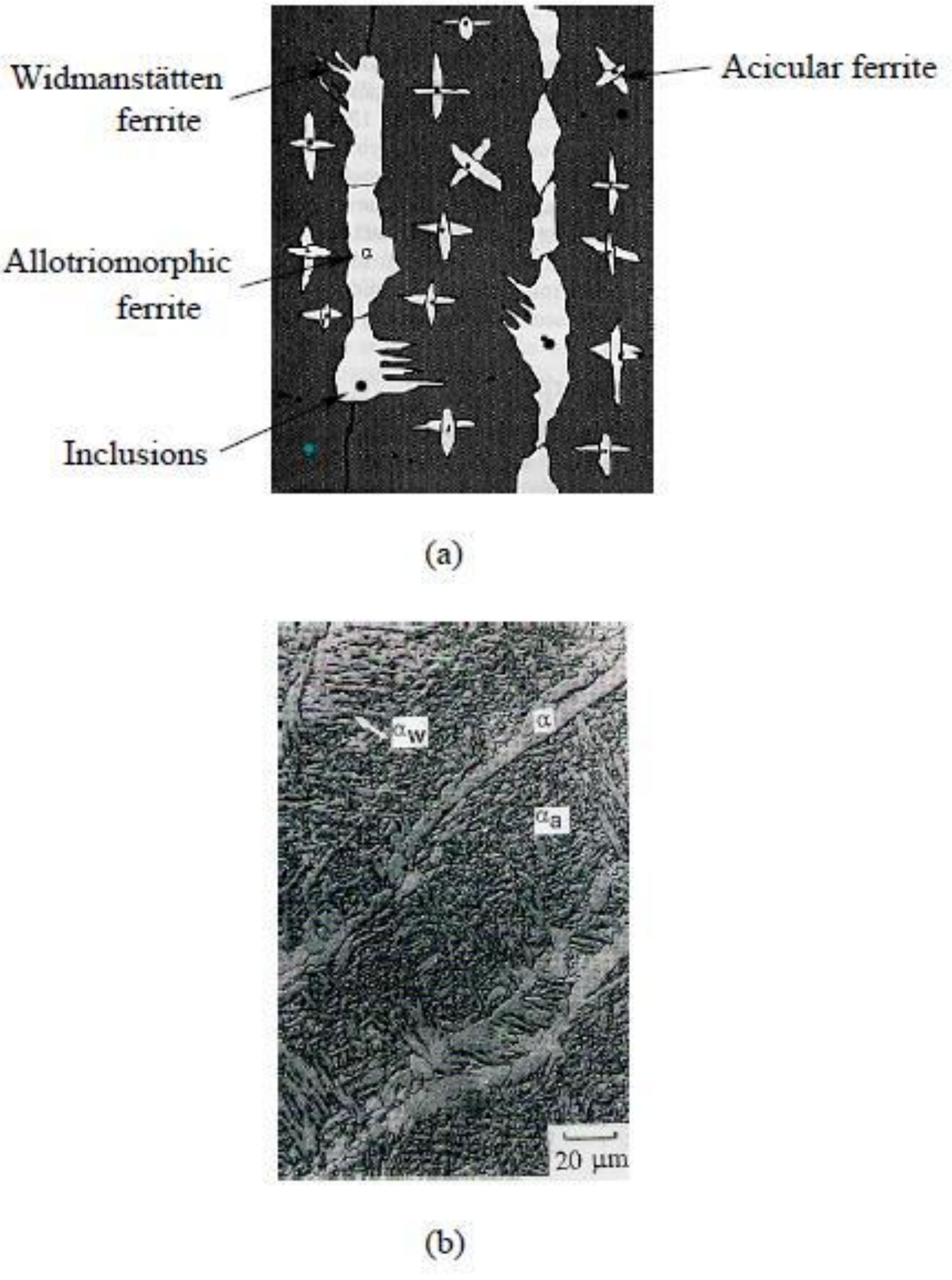


Figure 2.9 Schematic view of the various zones in a single pass weld metal.



(a)

(b)

Figure 2.10 a) Schematic diagram showing different constituents of the primary microstructure in the columnar austenite grains of a steel weld [13], b) scanning electron micrograph of the primary microstructure of a steel weld [14]. a- allotrimorphic ferrite, aw- Widmanstatten ferrite and aa- acicular ferrite.

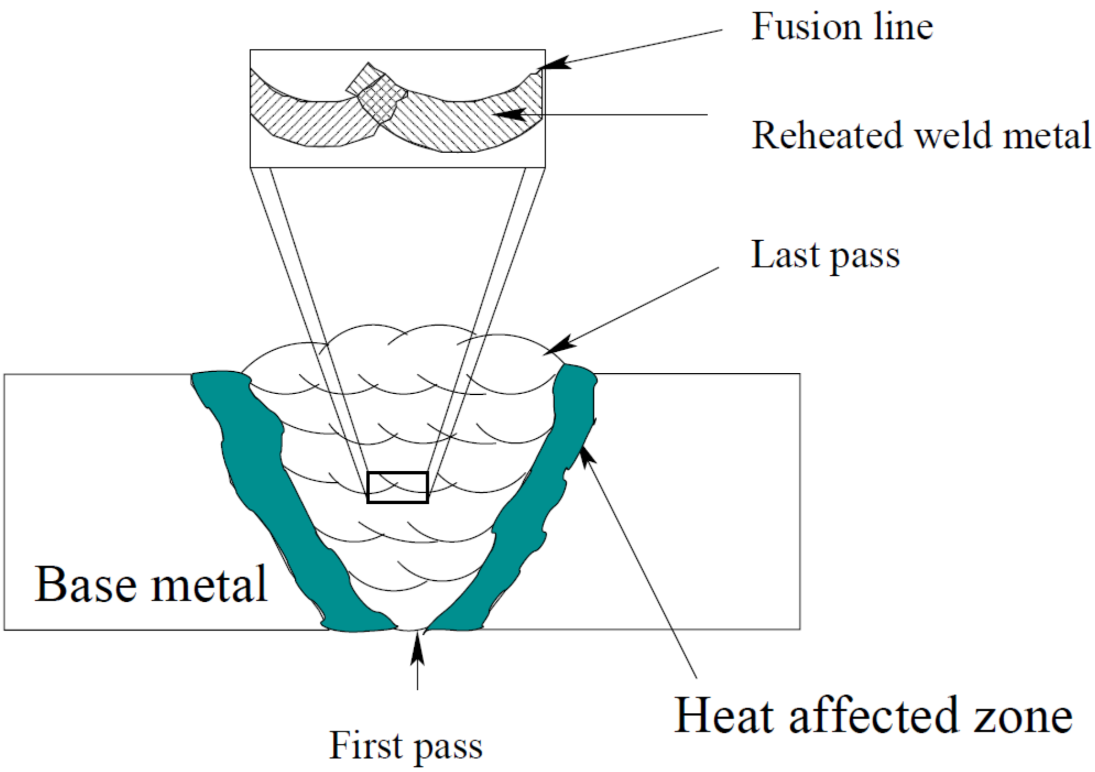


Figure 2.11 Various regions in a multilayer welding.

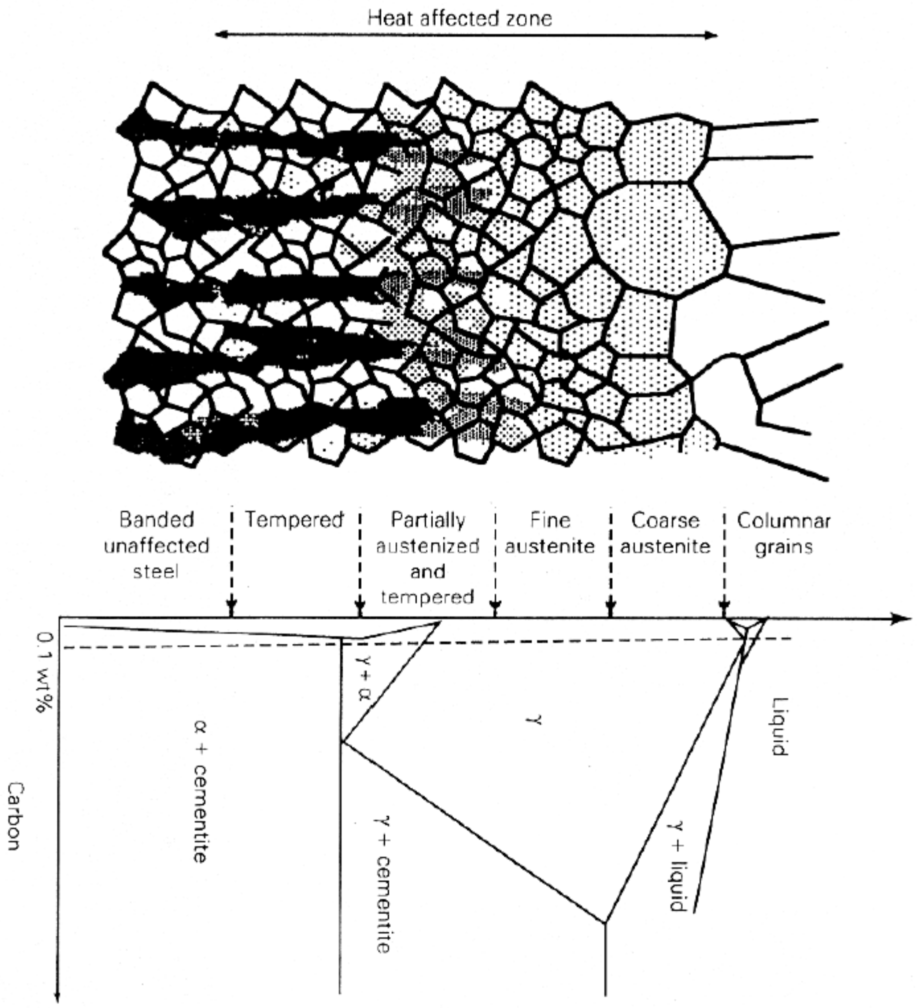


Figure 2.12 Microstructural variations in heat jected zone [14] The banded structure is a characteristic feature of segregated steels which have been rolled.

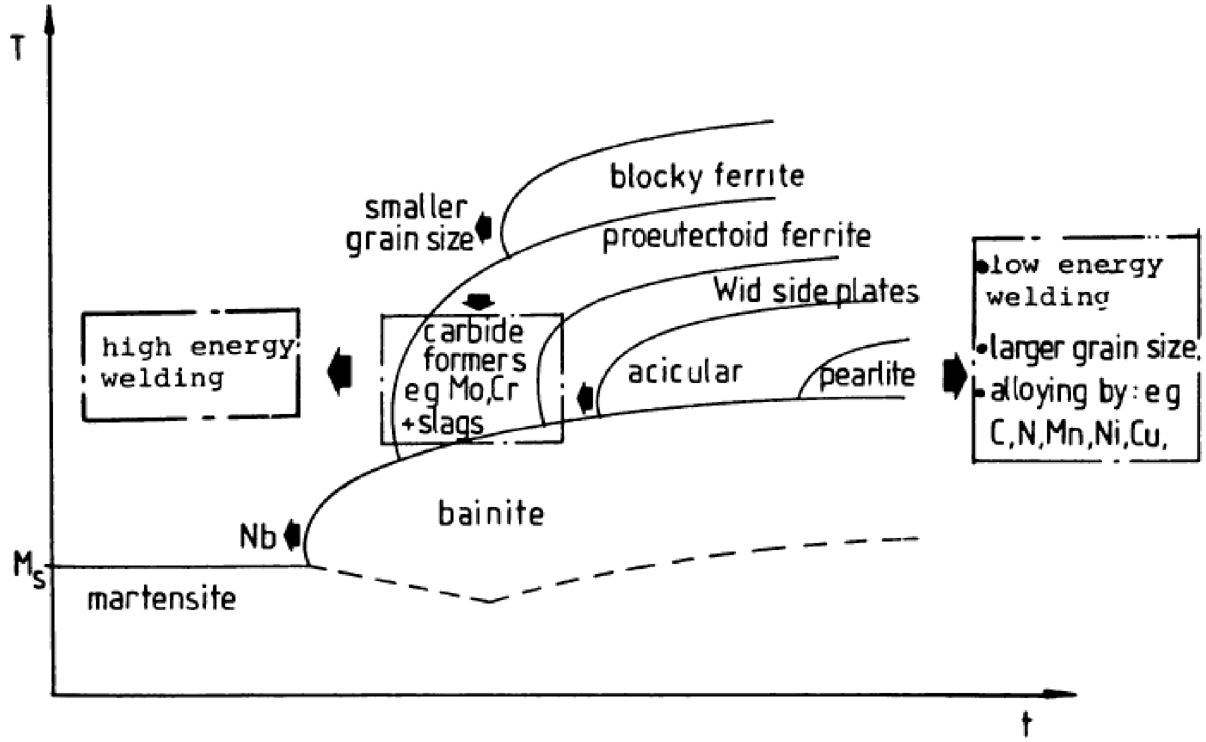


Figure 2.13. Schematic CCT diagram for steel weld metal, summanzmg the possible effect of microstructure and alloying on the transformation products for a given weld cooling time [10].

2.7 Strengthening Mechanism

Iron in its pure form is weak and can have a yield strength as low as 50 MPa [15]. The strength of pure body- centered cubic iron in a fully annealed condition decreases rapidly as the temperature is increased, Fig. 2.14. The strength is particularly sensitive to temperatures below -25 0C. In fact, it is this sensitivity to temperature which gives rise to the ductile-brittle transition. The cleavage strength of iron, is insensitive to temperature; at sufficiently low temperature it becomes less than the How stress, making iron brittle.

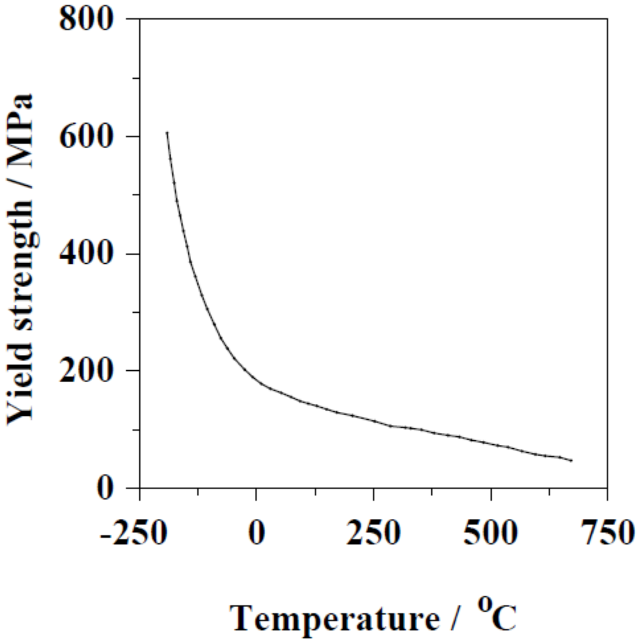


Figure 2.14 Temperature dependence of the yield strength of iron (gettered with titanium) at a plastic strain of 0.002 [16]. The strain rate is 2.5x I0- 4 s- I.

2.7.1 Grain Refinement

The refinement of grain size leads not only to an increase in the strength but also toughness [17]. Grain boundaries are formidable obstacles to the movement of dislocations. The dependence of the yield strength on grain size is given by the Hall-Petch relationship [18]:

$$\sigma_y = \sigma_i + k_y d^{-1/2} \tag{2.8}$$

where „d' grain diameter, sy is the yield stress, si is the friction stress opposing the movement of dislocation in the grains and Ky is a constant. The derivation of the Hall-Petch equation relies on the formation of a dislocation pile- up at a grain boundary, one which is large enough to trigger dislocation activity in an adjacent grain. Yield in a polycrystalline material is in this context defined as the transfer of slip across grains.

A larger grain is able to accommodate more dislocations in a pile- up, enabling a larger stress concentration at the boundary, thereby making it easier to promote slip in the nearby grain [19].

It is harder to propose a general mechanism by which grain refinement improves toughness. The argument for steels is that grain boundary cementite particles are finer when the grain size is small [19]. Fine particles are more difficult to crack and any resulting small cracks are difficult to propagate, thus leading to an improvement in toughness.

2.7.2 Solid Solution Strengthening

The most common method of increasing the hardness and strength of steels is by solid solution strengthening. The degree of hardening or softening due to dissolved elements depends crudely on the relative difference in atomic size relative to an iron atom [16]. Large atoms induce compressive stress fields whereas smaller atoms are associated with tensile fields in the matrix. These distortions interact with dislocation motion. Solid solution strengthening also depends on disturbances to the electronic structure, expressed in terms of the difference of the solute and host atom [17].

In steels the smaller atoms carbon and nitrogen occupy interstitial sites whereas elements like silicon, manganese are substitutional. The interstitial solute atoms cause an asymmetrical distortion of the ferrite lattice whereas the substitutional solute produce symmetrical distortions. Therefore the increase in strength of a-iron by interstitial carbon or nitrogen is much greater than that of any substitutional alloying element Fig. 2.15. Isotropic distortion can only interact with the hydrostatic stress fields of dislocations. Much greater interactions are possible with the tetragonal strains associated with the interstitial atoms in ferrite.

Fig. 2.16 shows that the strengthening due to substitutional solutes often goes through maximam as a function of temperature. In a few cases there is some softening in body centered cubic a- iron at low temperatures because the presence of foreign solute atom locally assists dislocations to overcome the large Peierls barrier to dislocation motion [4].

2.7.3 Precipitation Hardening

Small and uniformly distributed precipitates can be effective barriers to dislocation movement. Precipitation hardened steels strengthening are usually first heat treated in the austenite phase Held in order to dissolve solutes and then cooled rapidly to ambient temperatures to produce a supersaturated ferrite or martensitic transformation. Tempering then allows the excess solute to precipitate as carbides or nitrides, thereby strengthening the microstructure. In steels the strong carbide- forming elements titanium, vanadium, niobium, molybdenum, etc. are commonly used as the main precipitation strengthening elements. This mechanism is applied widely to increase the creep strength of power plant steels.

2.7.4 Post Weld Heat Treatment

During welding, there are regions created which are austerities and then cooled rapidly, producing brittle microstructures such as martensite. Tempering is frequently used to restore the toughness, by heat treating in a temperature regime where austenite cannot form. Thus, any excess carbon in solution is rejected to form carbides. In some cases the purpose of tempering is to induce the precipitation of alloy carbides. Power plant steels containing carbide forming elements such Cr, Mo, V, Nb, Ti, and W form stable carbides such as MX , M3 X, M2 X, M7 X3 , M23X6 and M6X (where M represents metal atoms, X represents interstitial atoms) on tempering at temperatures where there is sufficient mobility for the diffusion of substitutional atoms. This generally means temperatures above 500 ^{o}C. The precipitation of alloy carbides and the associated strengthening is often referred to as secondary hardening' [20]. Fig. 2.17. Shows the variety of carbides formed during tempering of water quenched 2 Cr-lMo steel from its austenitisation temperature.

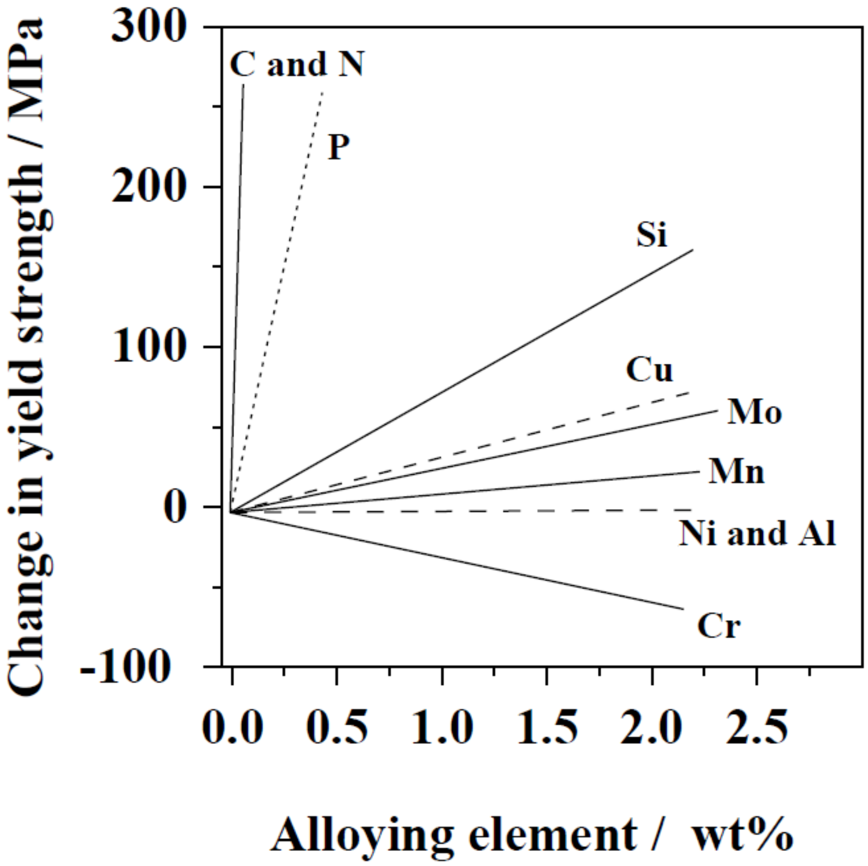


Figure 2.15 Contributions to the solid solution strengthening of ferrite [17].

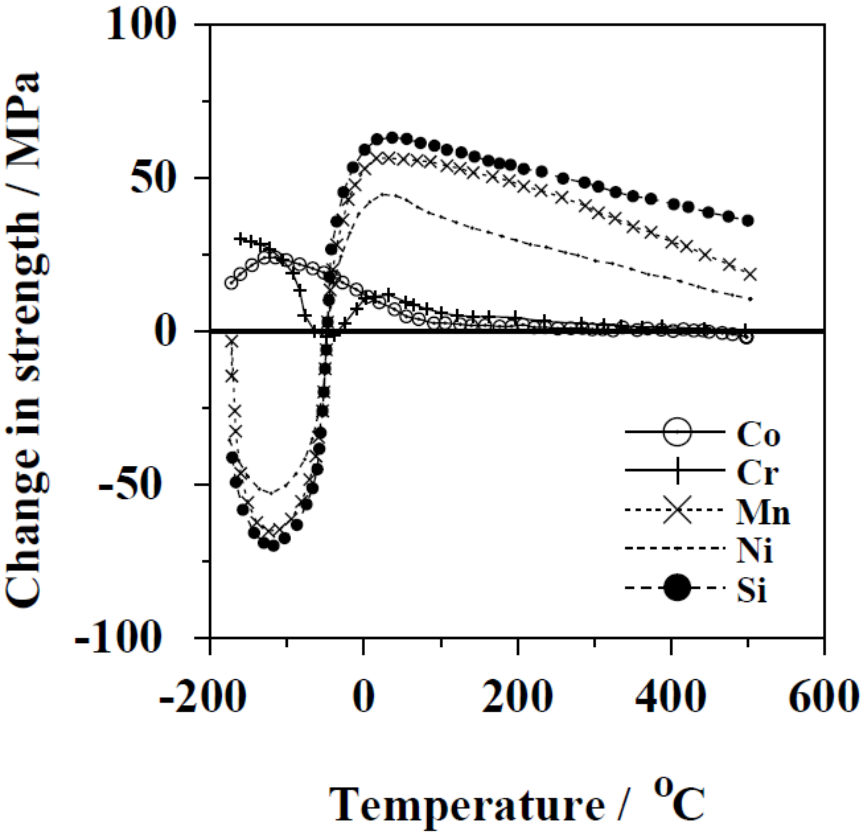


Figure 2.16 The effect of some substitutional solutes (3 at. %) on the yield strength of iron [16]. The strain rate is 2.5 x 10-^{4} s-^{1}.

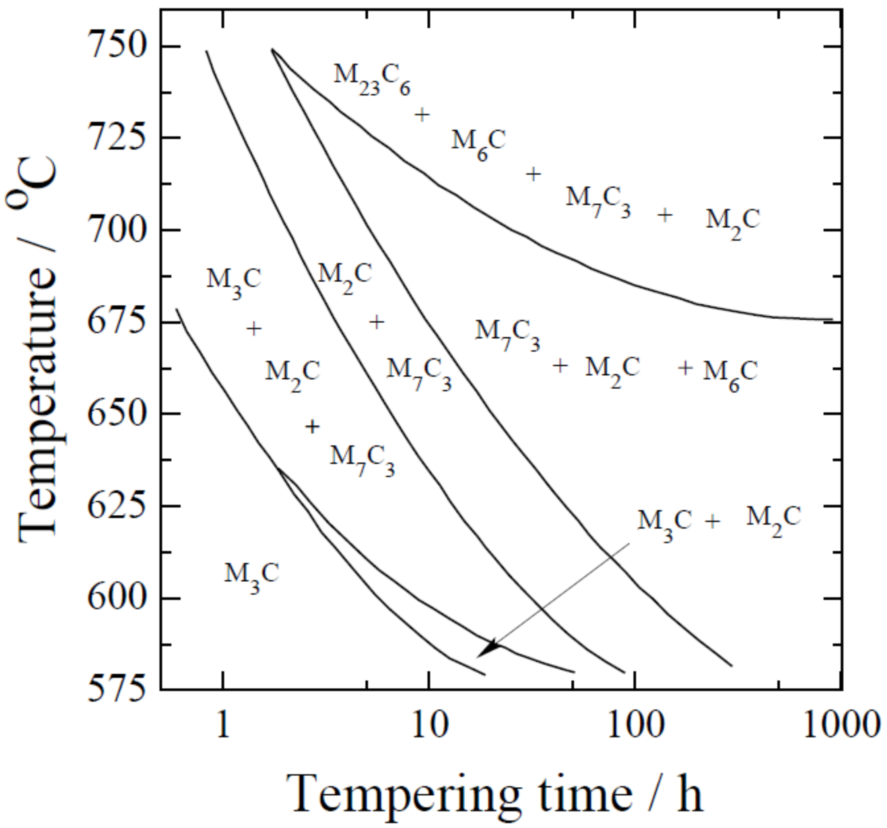


Figure 2.17 Carbide sequence in water quenched 2¼Cr-1Mo steel [21], where 'M' represents metallic elements.

2.8 Previous Weld Mechanical Property Models:

Weld metal models can in general be categorized into two classes, those which are empirical and others founded on physical metallurgy. The latter are more meaningful, but as will be seen later, they are generally over-simplified and deal only with simple properties rather than the range of properties important in engineering design.

1. Regression Models

   There have been numerous attempts to model weld metal mechanical properties by using linear regression analysis.

Table.2.2 Yield and Ultimate tensile strength (MPa) regression models of weld metals [1]. The alloying element concentrations are expressed in wt%.

Carbon-Manganese	$YS = 335 + 439C + 60Mn + 361\ (C.Mn)$ $UTS = 379 + 754C + 63Mn + 337\ (C.Mn)$
Silicon-Manganese	$YS = 293 + 91Mn + 228Si - 122Si^2$ $UTS = 365 + 89Mn + 169Si - 44Si^2$
Chromium-Manganese	$YS = 320 + 113Mn + 64Cr + 42\ (Mn.Cr)$ $UTS = 395 + 107Mn + 63Cr + 36\ (Mn.Cr)$
Nickel-Manganese	$YS = 332 + 99Mn + 9Ni + 21\ (Mn.Ni)$ $UTS = 401 + 102Mn + 16Cr\ 15\ (Mn.Ni)$

The strength of weld metal is frequently modeled as a function of chemical composition of weld metal, for cases where all the remaining variables associated with welding approximately constant. Equations like these are useful within the context of the experiments they represent. Naturally, the firm of the relationships used may not necessarily be justified in detail.

2. The Sugden-Bhadeshia Model

   Sugden and Bhadeshia tried to predict the strength of the as-deposited weld as a function of the chemical composition and microstructure [8]. The model is based on the assumption that the strength can be factorised into components; strength of pure iron, solid solution strengthening and strength due to microstructure, equation 1. The chosen microstructural constituents are allotriomorphic ferrite (α), Widmanstatten ferrite (α_w), and acicular ferrite (α_a) with the following assumptions:

2.1 The total strength ($\sigma_{y)}$ of as-welded deposit is assumed to be a linear combination of individual components:

$$\sigma_y = \sigma_{\mathrm{Fe}} + \sum_{i=1}^{n} \sigma_{SS,i} + \sigma_{\mathrm{Micro}} \tag{2.9}$$

where σ_{Fe} is the strength of fully annealed pure iron as a function of temperature and strain rate, σ_{ss1i} is the solid solution strengthening due to alloying element i and σ_{Micr} is strengthening due to weld microstructure.

The weld microstructure consists of allotriomorphic Ferrite (α), Wimanstetten ferrite (α_w) and acciular ferrite (α_a). The variation in grain sizes of α, $\alpha_{w,}$ and α_a are not taken in to account:

$$\sigma_{\mathrm{Micro}} = \sigma_\alpha \ V_\alpha + \sigma_a \ V_a + \sigma_w \ V_w \tag{2.10}$$

where σα, σa and σw denote the contributions from 100% allotriomorphic ferrite, Widmanstetten ferrite an accicular ferrite respectively, and Vα, Va and Vw are their corresponding volume fraction.

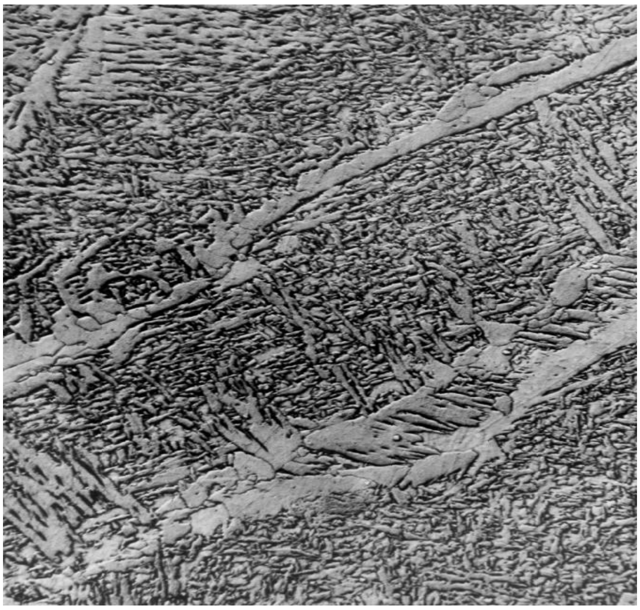


Figure.2.18 The weld microstructure consists of allotriomorphic ferrite (α), Wimanstetten ferrite (α_w) and acciular ferrite (α_a).

Nitrogen is assumed to be in solid solution and any Strain ageing effects in the as-welded microstructure are assumed to be negligible.The solid solution strengthening (σss) is expressed as the sum of the contributions from each solute:

$$\sigma_{SS} = a \text{ Mn wt\%} + b \text{ Si wt\%} + ... \tag{2.11}$$

where the coefficients a, b, .. are functions of temperature, defining the role of the respective alloying elements. The values for these coefficients are taken from the published experimental data which are based on studies in which solid solution strengthening is studied in isolation.

An alloying element naturally influences more than just solid solution effect. However, the other consequences are included in the analysis via incorporation of microstructure. The authors were able to estimate the strength of individual microstructures (σα, σa and σw) by studying three welds which are made with identical welding conditions [8]. The chemical compositions were adjusted to give different fractions of microstructure in order to deduce the strengthening due to each microstructure (α, α_a and α_w). The final form of developed equation is:

$$\sigma_y = \sigma_{\mathrm{Fe}} + \sigma_{SS} + 27V_\alpha + 402V_a + 486V_w \quad (\mathrm{MPa})$$

(2.12)

where σ_{Fe} and σ_{ss} can be obtained from referred published literature [22].

Although the Sugden-Bhadeshia model has more physical meaning when compared with the empirical equation presented in Table.2.2., the model still has linear approximations which are not justified in detail. It is resticted to structural steel welds which have simple, untempered microstructures bainite and martensite are excluded from the analysis,as is precipitation hardening. Young and Bhadeshia have developed the work for microstructures which are mixtures of bainite and martensite but this model has yet to be applied to weld metal microsructures. The model is nevertheless discussed below because it is interesting.

3. The Young-Bhadeshia Model
   The Young-Bhadeshia strength model for high-strength steels [4] considered microstructures which are mixtures of martensite and bainite;

$$\sigma = \sigma_{\mathrm{Fe}} + \sum_i \sigma_{SS,i} + \sigma_C + K_L(\overline{L})^{-1} + K_D\rho_D^{0.5} + K_p\Delta^{-1}$$

(2.13)

where $K_{L,}$ K_D and K_p are constants, σ_c is the solid solution strengthening due to carbon, L is a measure of the ferrite plate width, ρ_D is the dislocation density and Δ is the distance between any carbide particles. The other terms have their usual meanings.

The Young and Bhadeshia model can be applied to estimate the strength of bainite and martensite welds by using rule of mixtures. Even though the model had considered the microstructural influence the model still built on the some of the assumptions made in Sugden and Bhadeshia model like linear summation effect of solid solution strengthening.

It appears from the literature reviewed that the failure of the previous work [1,2,3] to create models with wide applicability comes largely from constraints due to the linear or pseudolinear regression methods used, with poor error assessments and most importantly from very limited variables and data considered in the analysis.

2.9 NEURAL NETWORKS OVERVIEW

Neural networks are powerful and demonstrably useful tools for solving practical problems. Noteworthy applications include medical and machine diagnosis, machine control, credit risk and financial prediction, weather prediction, and even prediction of outcomes in such exotic Endeavors as horse racing. Neural networks have seen an explosion of interest over the past few years.

Even problems that are considered intractable by conventional methods have yielded to neural networks analysis. Because conventional linear statistical models have well known optimization strategies, linear modeling has been most commonly used modeling technique in many problem domains for a very long time; but where the linear approximation was not valid(often the case), the models suffered accordingly.

Advantages of neural network techniques over conventional techniques include the ability to address highly nonlinear relationships, independence from assumptions about the distribution of input or output variables, and the ability to address either continuous or categorical data as either inputs or outputs.

Neural networks are also intuitively appealing, based as they are on crude low-level models of biological systems. As in biological systems, neural networks simply learn by example, but in the case of neural networks, the neural network user provides representative data and trains the neural networks to learn the structure of the data.

2.10 Bayesian neural networks

An artificial neural network (ANN) is basically a method for fitting a curve to a number of points in data space[6]. More technically, it is a parameterized non- linear model which can be used to perform regression, in which a flexible, non-linear function is fitted to experimental data. The term “artificial” is used to indicate that these networks are computer programs, rather than “real” neural networks such as the human brain. The details of the operation and construction of neural networks have been reviewed elsewhere [6], but it is useful to summarise the main features here.

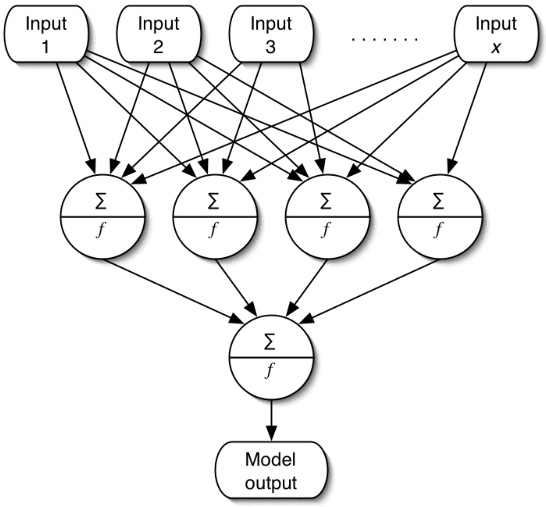


Figure 2.19 A schematic diagram of a three-layer feed-forward network. The model‟s complexity is controlled by the number of neurons in the second layer, known as hidden units.

It has been shown that a sufficiently complex three-layer network of the form described below can imitate any complex function [6]. The network is thus able to respond flexibly to the demands made by the data, capturing any non-linear interactions between the parameters.

Such a three-layer feed-forward network, of the type commonly used for material property applications, is shown in Figure 2.19. The first layer consists of the inputs to

the network. The second layer consists of a number of neurons – non-linear operators whose arguments are provided by the first layer in the network. The activation function for these neurons, h_i, can be any non-linear, continuous and differentiable function – tanh has been used in this work (Equation 2.14). The overall output function, y, can again be any function, and is commonly linear. The neuron activation function for a neuron i is given by

$$h_i = \tanh\left(\sum_j w_{ij}^{(1)} x_j + \theta_i^{(1)}\right) \tag{2.14}$$

and the output weighting function is

$$y = \sum_i w_i^{(2)} h_i + \theta^{(2)} \tag{2.15}$$

The x_j are the inputs, and w the weights which define the network. $^{(1)}$ and $^{(2)}$ denote weights and biases in the hidden layer and in the output layer, respectively. The aim of training a network is to find the optimum set of values for w. The parameters θ are known as biases, and are treated internally as weights associated with a constant input set to unity.

In order to simplify the weightings, inputs are normalised within a range of ±0.5.

The normalisation function is

$$x_j = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \qquad (2.16)$$

where x is the un-normalised input, x_{min} and x_{max} are the minimum and maximum values in the database for that input, and x_j is the normalised value. The network is therefore not constrained to a particular range of outputs (for example, positive outputs only) and so the target must be chosen with care to avoid unphysical model outputs. For example, for a property which cannot be less than zero such as the yield stress σ_y, ln(σ_y) could be used instead as the network training target.

The complexity of such network models scales with the number of "hidden" units.

Despite the terminology and the common view of a neural network as a "black box", the weightings can in fact be examined although they are difficult to interpret directly, being complex nested tanh functions. The easiest way to identify the interactions in a model is to use it to make predictions and see the behaviour which emerges from various combinations of inputs.

Because of the inherent flexibility of an ANN, there is the possibility of overfitting the model. Training a network therefore involves finding a set of weights and biases which minimize an objective function, which balances complexity and accuracy, typically

$$M(w) = \alpha E_w + \beta E_D \qquad (2.17)$$

in which E_W is a regulariser; its function is to force the network to use small weights and limit the number of hidden units and is given by

$$E_w = \frac{1}{2} \sum_{ij} w_{ij}^2 \qquad (2.18)$$

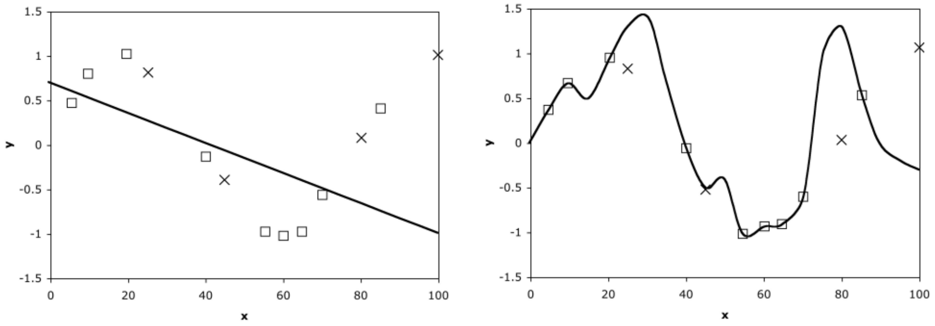
and E_D is the overall error between target output values and network output values, given by

$$E_D = \frac{1}{2}\sum_k \left(t^{(k)} - y^{(k)}\right)^2 \tag{2.19}$$

where $t^{(k)}$ is the set of targets for the set of inputs $x^{(k)}$, while $y^{(k)}$ is the set of corresponding network outputs. α and β are control parameters which influence the balance between a simple but inaccurate model, and an overcomplex, also inaccurate model (Figure 2.20). MacKay"s algorithm allows the inference of these parameters from the data, permitting automatic control of the model complexity [6].

To accomplish the training, the data are randomly split into two sets, a training set and a test set. The model is trained on the training set, and then its ability to generalise is compared against the test set of data. Figure 2.21 shows how increasing complexity continuously lowers the training error (the mismatch between model predictions and the training dataset), while the test error (the mismatch between model predictions and the test dataset) has a minimum. At greater complexities, overfitting causes the test error to increase with increasing numbers of hidden units. The ultimate purpose of training a model is to minimise this error, both against the input dataset and against unseen data from future experiments.

For these models, the fitting method is based on a Bayesian approach and treats training as an inference problem, allowing estimates to be made of the uncertainty



(a) (b)

Figure 2.20 Under- and over-fitting. A set of noisy data points (hollow boxes) has been fitted by (a) linear regression and (b) an overly complex function. In the first case the fit clearly does not represent the data, and in the second case the fit overlies the training data perfectly but generalises poorly to new points (crosses) of the model fit (Figure 2.22).

Rather than trying to identify one best set of weights, the algorithm infers a probability distribution for the weights from the data presented. In this context, the performances of different models are best evaluated using the log predictive error (LPE) rather than the test error. This error penalises wild predictions to a lesser extent when they are accompanied by appropriately large error bars and is defined by

$$\mathrm{LPE} = \frac{1}{2}\sum_{k}\left[\frac{\left(t^{(k)}-y^{(k)}\right)^{2}}{\left(\sigma_{y}^{(k)}\right)^{2}} + \log\left(2\pi\left(\sigma_{y}^{(k)}\right)^{2}\right)\right] \tag{2.20}$$

where t and y are as defined above, and $\sigma^{(k)}_{y}$ is related to the uncertainty of fitting for the set of inputs $x^{(k)}$. It should be pointed out that, for computational purposes, the training software (BigBack5[1]) actually uses an inverse version of this function that increases with increasing accuracy.

Of course, models with different number of hidden units and different initial guesses for the distribution of the weights, the prior will give different predictions. Optimum predictions can often be made by using more than one network. This is referred to as a committee. The prediction y of a committeee of networks is the mean prediction of its members, and the associated uncertainty is used to give the corresponding prediction $y^{(l)}$.

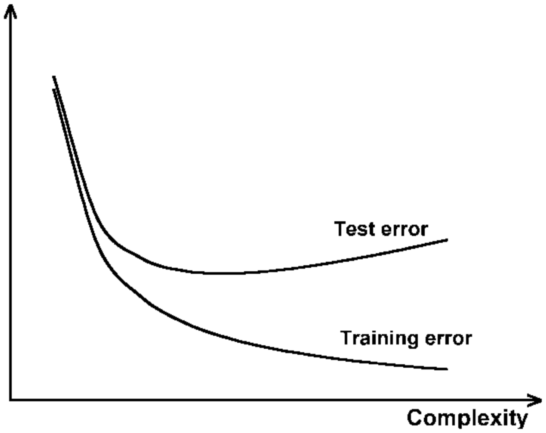


Figure 2.21 Comparison of error on training and testing sets as a function of network complexity, illustrating the problem of over complex models as in Figure 2.20.

$$\sigma^2 = \frac{1}{L}\sum_l \sigma_y^{(l)2} + \frac{1}{L}\sum_l \left(y^{(l)} - \overline{y}\right)^2 \tag{2.21}$$

where L is the number of networks in the committee and the exponent (l) refers to the model During training, it is usual to compare the performances of increasingly large committees on the testing set of data. Usually, the error is minimised by using more than one model in the committee. The selected models are then retrained on the entire database.

A further output from the training software is an indicator of the network-perceived significance of each input. The measure provided by BigBack5 is a function of the values of the regularisation constants for the weights associated with an input, σ_W . This measure is similar to a partial correlation coefficient in that it represents the amount of variation in the output that can be attributed by any particular input.

To determine the sensitivity of the model to individual input parameters, on the other hand, predictions must be made varying one parameter only whilst keeping all the others constant. In some cases where an input is a function of one or more other inputs (for example, both temperature T and an Arrhenius function exp $(l)^2$ could be inputs to the network) varying only one of these parameters may not be physically meaningful.

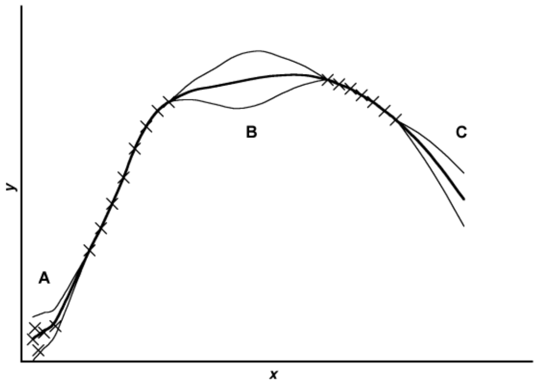


Figure 2.22 Schematic illustration of the uncertainty in defining a fitting function in regions where data are sparse (B) or noisy (A). The thinner lines represent error bounds due to uncertainties in determining the weights. Note that, outside the range of data, the extrapolation is increasingly uncertain (C). Areas of high uncertainty will provide the most informative new experiments.

The nature of the ANN structure (and these outputs) allows the "testing" of various physical models – input parameters based on those models can be included in the training data, and those parameters which are not useful in explaining the output will have much lower significances than those which are useful

2.11 Theory of Multilayer perceptron (MLP). Radial Basis function (RBF). Generalized Regression neural networks (GRNN)

2.11.1 KEY ELEMENTS OF NEURAL NETWORKS

Let"s assume that we have collected a large number of data cases consisting of three continuous and a single categorical variable with three classes, and that we want to build a model that will be able to predict the association class using the three continuous variables as inputs. Instead of invoking a linear modeling approach such as discriminant analysis in which we have to make assumptions about the underlying distribution of the input and output variables or error terms, we could employ a simple type of neural network.

The figure below is a schematic representation of a simple neural network appropriate for the problem described above. We can use this example of *a Multilayer perceptron (MLP*) network to illustrate many of the elements of neural networks. Most of the elements will be formed in other network architectures supported by SNN. [23]

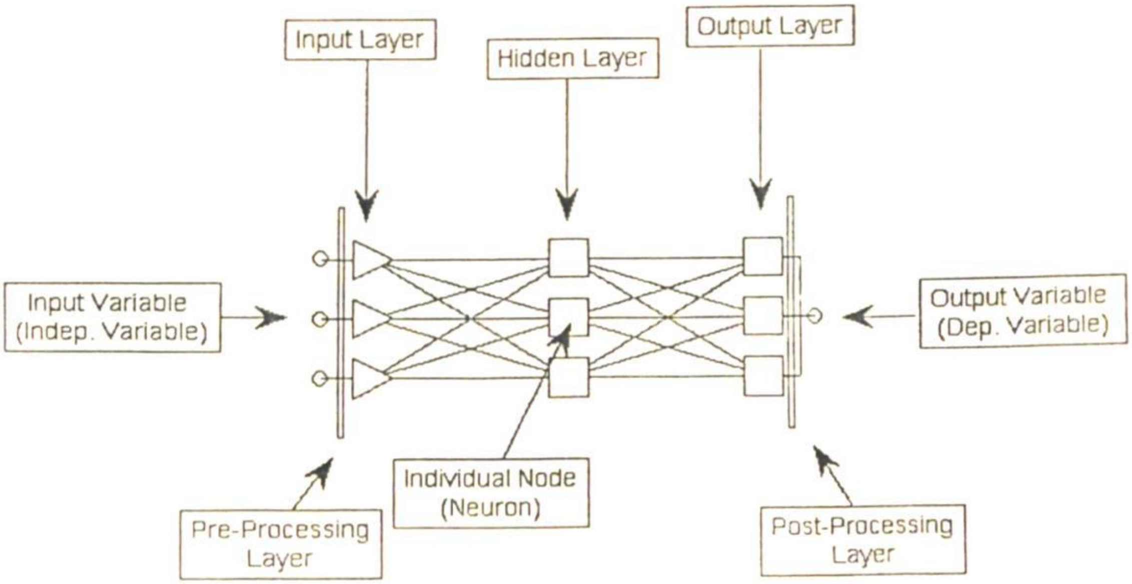


Figure 2.23 A schematic representation of a simple neural network with the elements.

The network can be thought of as consisting of consecutive layers progressing from inputs (left) to outputs (right). The individual neurons (nodes/units) within the layers of this common neural network type consist of mathematical constructs that, when properly exposed to historical data (trained), will internally adjust to a configuration that will allow the network accept input values, process them through the various layers of calculations, and finally, produce a predicted value (or classification) as the last layer‟s output.

The raw data that are contained in the three input or independent variables (left side of the figure) need to be transformed to a range of values that can be used in the neural network. In STATISTICA Neural networks (SNN), a separate conversion or pre-processing layer that is a part of the network architecture performs the transformation. In the case of the continuous input variables employed in this example network, the raw values simply need to be rescaled. (For categorical input variables, the pre-processing/conversion function would map the categorical values into a numerical form.)

The transformed values are fed to the input layer of neurons (also referred to as nodes or units). The input layer contains one unit for each of the input variables. (In the case of categorical inputs, one input unit will be created for each class or category a categorical input variable.)

Each unit in the input layer is connected to each unit in the hidden layer. The hidden layer is the layer that controls the amount of complexity that can be represented in the relationship between input and output variables. The larger the number of units in the hidden layer, the more complex/nonlinear is the relationship that can be represented. If no hidden layer were present then the neural network would describe a linear relationship between the input and output variables. Some network types used to model extremely complex relationships may contain two hidden layers.

In turn, each unit in the hidden layer is connected to each unit in the output layer. The output layer in this instance contains three units, one for each class of the categorical output or dependent variable. Finally, the outputs from the output layer of units are passed to a post-processing/conversion layer for mapping of the numerical information contained in the output layer units to classes of the categorical output variable.

The schematic in the figure 2.24 below depicts what goes on at each within the neural network. We can assume that the unit we see here is the topmost unit in the hidden layer shown in the previous figure. This unit like every individual unit in the network receives output values from all of the units in the layer preceding it. (In SNN all networks are of this feed-forward type; no networks are employed that feed information back to previous layers). The unit might be thought of as consisting of two parts, an input half(on the left) that accepts outputs(X) from the previous layer and an output half(on the right) that modifies information received from the left half and passes it on to units of the succeeding layer.

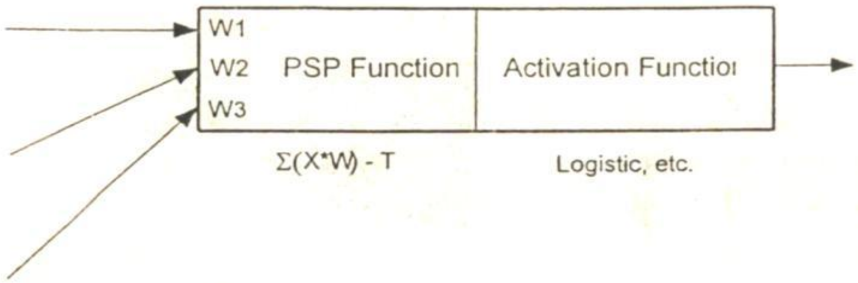


Figure 2.24 shows the functions in Neural Networks.

Each individual input coming into the unit is multiplied by a weight (W) value that the unit retains for that specific input. The sum of the products of the input values and their corresponding weights is compared with a threshold value (T) that is also retained by the unit.

This threshold value is also known under the term bias. The comparison of the weighted inputs with threshold value is (in this case) a linear equation and is commonly referred to as a *post-synaptic potential (PSP) function*. The resulting value (called the activation level) is passed on to the second half of the unit. When the network is created, random small weight values are assigned to the inputs arriving from the units in the previous layer and random values are also assigned to the thresholds.

The second half of the unit is commonly referred to as an activation function. In the hidden layers(s) this is usually an S-shaped (or sigmoid) function that accepts the value of the activation level received and rescales it to a value within a defined range that the subsequent units in the network can accept. In input and output layer units, the activation function is typically linear.

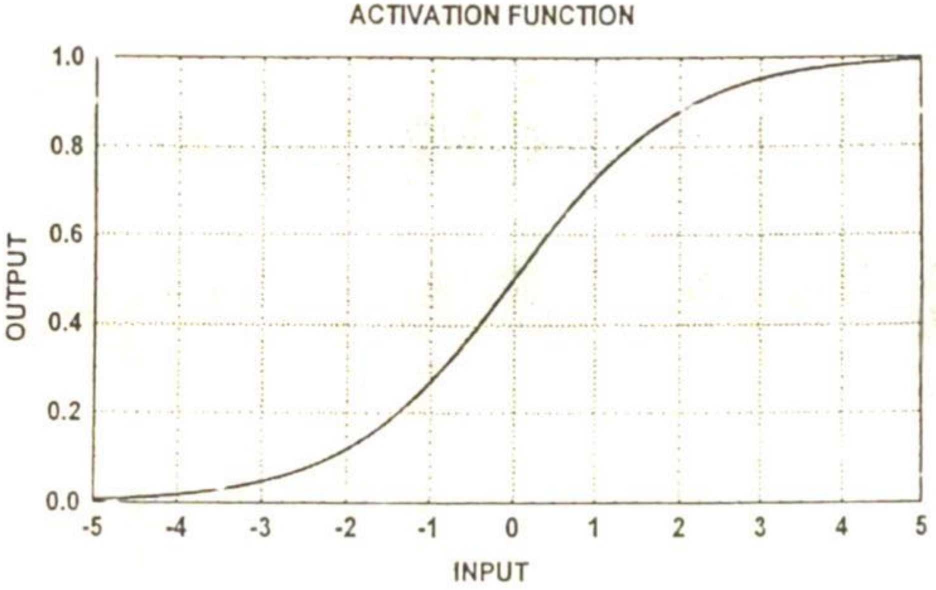


Figure 2.25 Activation function in Neural Network

By exposing the network repeatedly to historical data (i.e., data in which the input data values and the outcomes are known), the weights (Ws) and thresholds (Ts) of the PSP function are adjusted using special training algorithms until the network. Typically a subset of the data (Training data) is presented to the network in several or even hundreds of iterations. Each presentation of the training data to the network for adjustment of Ws and Ts is referred to as an epoch. The procedure continues until the overall error function has been sufficiently minimized, as discussed in the next paragraph.

The overall error is also computed for a second subset of the data: selection data (sometimes referred to as verification or validation data). This data takes no part in the adjustment of Ts and Ws during training, but the network"s performance is continually checked against this subset as training continues. If the error for the selection data stops decreasing or starts to increase, the training is stopped. Use of a selection subset of data is important, because with unlimited training, the neural network usually starts "overlearning" the training data. Given no restrictions on training, a neural network may describe the training data almost perfectly but may generalize very poorly to new data. The use of selection subset to shut down training at a point when generalization potential is best is a critical consideration in training neural networks.

In most cases it is even advisable to create a third subset of Test data to serve as an additional independent check on the generalization capabilities of the neural network.

2.12 NETWORK TYPES

STATISTICA Neural networks provides a wide selection of network architectures and respective training algorithms. Each architecture has advantages and disadvantages and is capable of performing certain tasks. Without going into computational details, the following overview describes the different network architectures, provides insights into different activation and error functions used, and lists the type of training algorithms available in STATISTICA Neural networks that can be used for training these networks. [23]

2.12.1 Multilayer perceptron(MLP)

This is perhaps the most popular network architecture is use today and discussed at length in most neural network textbooks. Multilayer perceptrons use a linear PSP function (i.e., they perform a weighted sum of their inputs), and a (usually) nonlinear activation function. A three-layer MLP (i.e., with one hidden layer) has the capability to model problems of almost any degree of complexity. STATISTICA Neural networks provides options for creating MLP networks with up to three hidden layers.

The standard activation function for MLPs is the logistic function; STATISTICA Neural networks selects this for all layers by default. The hyperbolic function (tanh), which is a symmetric version of the logistic function, can produce better performance than the logistic function in many cases. If an MLP network is being used in regression problems, performance can often be improved by giving units in the output layer the linear activation function. This allows extrapolation beyond the training data in addition to interpolation. If an MLP network is being used in a single-output classification problem, the output layer error function can be changed to *Entropy* (single) for improving performance. If an MLP network is being used in a multiple –output classification problems, the output layer activation function can be set to *softmax*, in which case the network"s error function needs to be set to *Entropy*(multiple).

MLP networks can be trained by any of the following algorithms. *Conjugate Gradient Descent, Quasi-newton, Levenberg-Marquardt, Back Propagation, Quick Propagation, or Delta-bar-Delta.* MLP networks are relatively compact, and widely applicable; however, the training process can be protracted, and they are prone to meaningless extrapolation if given highly novel data.

2.12.2 Radial Basis function(RBF)

Radial basis function networks have an input layer, a hidden layer of radial units and an output layer of linear units. They are described in most good neutral network textbooks.

Typically, the radial layer has exponential activation functions and the output layer, a linear activation functions. Radial basis function networks are trained in three stages:

1. *Center-assignment.* The centers stored in the radial hidden layer are optimized first; typically using unsupervised training techniques. Centers can be assigned by a number of algorithms: by *sub-sampling, K-means, Kohonen training, or Learned vector quantization.*These algorithms place centers to reflect clustering.
2. *Deviation assignment.* The spread of the data is reflected in the radial deviations (stored in the threshold). Deviations can be assigned by a number of algorithms*(Explicit, Isotropic, K-nearest neighbor).*
3. *Linear optimization.* Once centers have been assigned, the linear output layer is usually optimized using the pseudo-inverse technique, as this is quick, and guaranteed to minimize the error if the deviation are not too small. However, you may also use *Conjugate Gradient Descent, Quasi-Newton, Back Propagation or Delta-Bar-Delta* training, if you wish. Alternatively, an RBF used for classification may use cross-entropy error functions and logistic or softmax activation functions. In this case, iterative training algorithms must again be used.

Radial Basis function networks train relatively quickly and do not extrapolate too far from known data; however, they tend to be larger than MLPs and therefore execute more slowly.

2.12.3 Generalized Regression neural networks(GRNN)

There are two types of Bayesian networks: probabilistic neural networks(PNN) which are used only for classification tasks and Generalized Regression neural networks(GRNN) which are only used for regression tasks.

Generalized Regression networks have exactly four layers: input, a layer of radial centers, a layer of regression units, and output. The radial layer units represent the centers of clusters of known training data. This layer must be trained by a clustering algorithm such as *Sub-Sampling, K-means, or Kohonen training*. The layer is typically large, but not necessarily as large as the number of training cases. The regression layer, which contains linear units, must have exactly one unit more than the output layer. There are two types of units: type A units calculating the conditional regression for each output variable, with the single type B unit calculating the probability density. The output layer performs a specialized functions: each unit simply divides the output of the associated type A unit by that of the type B unit, in the previous layer. A special PSP function (Division) is provided for this purpose.

Bayesian networks (GRNN) train extremely quickly(almost instantly), but are typically very large and therefore execute slowly.

2.13 TYPICAL PROBLEM-SOLVING APPROACH

It is important to keep in mind that analyzing data with neural networks presents a “black box” approach. One typically not interested in the model parameters (e.g., the individual network weights) or their significance in evaluating the goodness-of-fit of the model, but rather using the model to solve a practical application (e.g., predict future values). [23]

Using neural network analysis to solve a problem generally involves a number of steps:

1. **Accumulating a set of representative data from the problem domain.** A network‟s solution will only be as good as the data on which it is trained. If those data are unrepresentative or skewed in any way, the network‟s solution will suffer and not generalize well to the population.
2. **Preparing the data for the analysis.** Data to be inputted to a network often needs to be prepared (e.g., by removing outliers and re-scaling of the original data or converting categorical data to a nominal variable). Sometimes this is as simple as transforming the data so that its values fall within an acceptable range.
3. **Trying a variety of network types and sizes.** Often the best network type and size for a problem is not known. The size of the network, for some network types, is related to the complexity of the problem, with more complex problems demanding larger networks for an appropriate solution.
4. **Making sure that the network generalizes well.** A successful network solution is one that generalizes well to the population. When trying different network types, it is useful to reserve a set of data to test the network to make sure that is has not learned the peculiarities of the training data at the expense of generalizing well(i.e., over-fitting).
5. **Running the network on new data.** Once an acceptable network solution is settled upon, the network can be saved. When new data are encountered, predictions can be made by opening the network and executing it on the new data.

2.14 Neural Network and Genetic Algorithm Modeling

2.14.1 Genetic algorithm

The genetic algorithm is a model of machine learning based on the mechanism of natural selection and natural genetics [24], [25]. This is done by a random creation of a population of individuals, represented by chromosomes. This individuals are evaluated and undergo a process of evolution which start with a natural selection, inspired by Darwin's theory of evolution: the best individuals of a population are selected. Then a biological process occurs: some recombinations, as crossover and mutation, are made in order to create a new generation of individuals with the hope that this new one is better. The genetic algorithm is stopped when a target value is reached.

Genetic algorithms are viewed as optimization tools and allow solving problems for which an extremum solution is searched.

2.14.2 Process

We are looking for an input set (x1 ,x2 , . .. ,xj) which will give a desired output y. The genetic algorithms are based on biological theory and so, we can assimilate this set as a chromosome. Each input xi is the equivalent of a gene. In a first time, a population is randomly generated. This population contains 'rt chromosomes (x1 ,x2 ,. . ",xj). These inputs are entered in a model, created by a neural network, and we obtain the output y and the associated error for each chromosome. The chromosomes are then ranked according to their fitness. The best chromosomes are selected and subjected to operations, as well in the biological system, as crossover or mutation. In this way, we obtain the second generation. This new sets of chromosomes are then still evaluated and undergo selection, crossover and mutation mechanisms, as previously. We obtain our third generation and this process takes place for many generations until the fitness value is reached (Fig. 2.26).

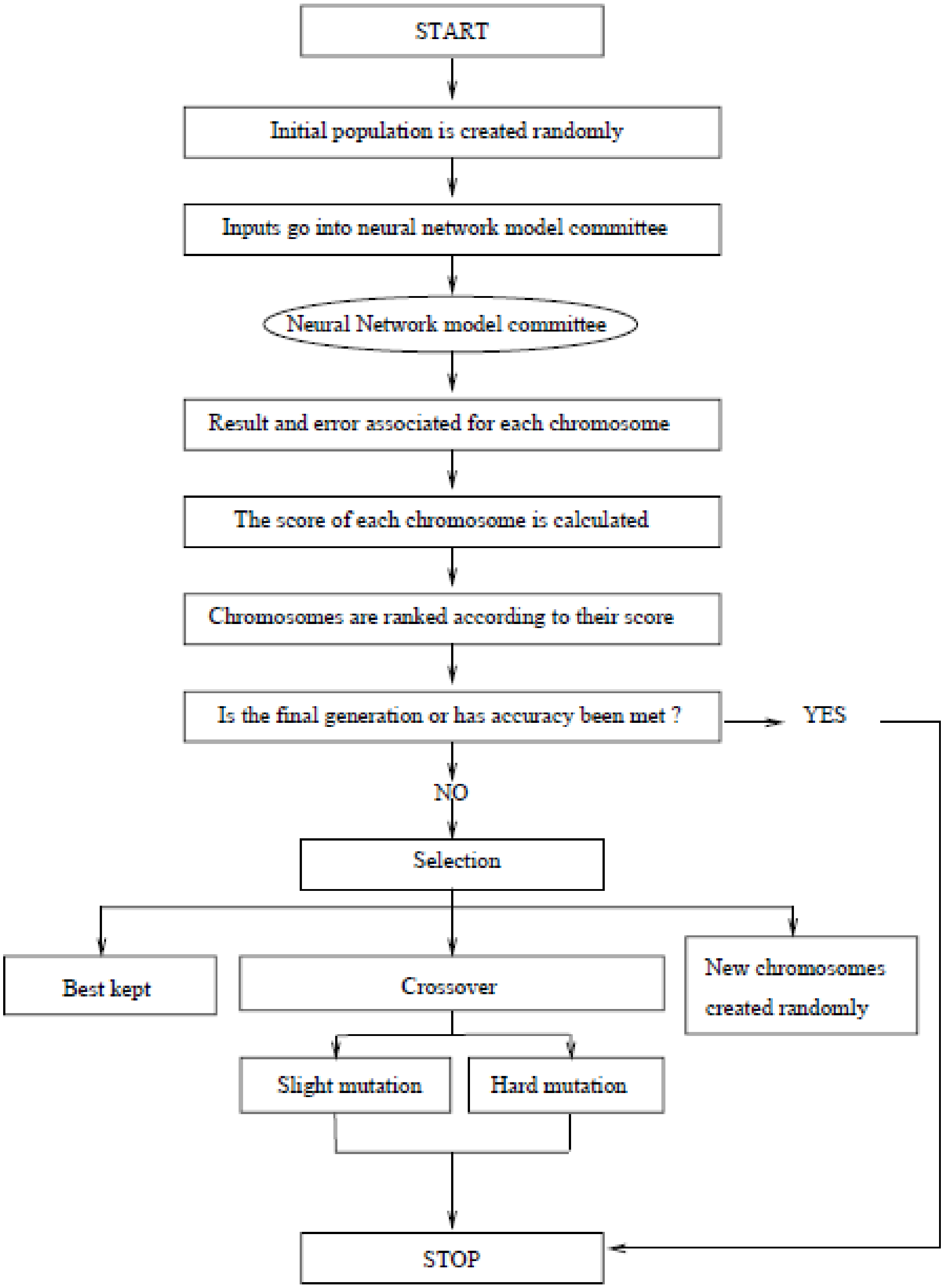

START
Initial population is created randomly
Inputs go into neural network model committee
Neural Network model committee
Result and error associated for each chromosome
The score of each chromosome is calculated
Chromosomes are ranked according to their score
Is the final generation or has accuracy been met ?
YES
NO
Selection
Best kept
Crossover
New chromosomes created randomly
Slight mutation
Hard mutation
STOP

Figure 2.26 The process of the genetic algorithm [Source Code: Appendix-B]

Operators

Fitness functions. The evaluation of a chromosome has to take into account the result and the error of the committee of models given by the neural network. Each model i created by the neural network gives a result y(i) and the associated error(s). The average prediction of a committee of L models is:

$$p = \frac{1}{L} \sum_{i} y^{(i)} \tag{2.22}$$

The standard deviation error(s) of p is as follows:

$$\sigma^2 = \frac{1}{L} \sum_{i} \sigma_y^{(i)^2} + \frac{1}{L} \sum_{i} (t - p)^2 \tag{2.23}$$

Where t is the desired output. The score of a chromosome could be s. however, in order to have a better score for the better chromosome, we invert the error and the fitness f is defined as follows:

$$f = \frac{1}{\sigma} \tag{2.24}$$

Selection

After being ranked according to their fitness, the chromosome undergo a process of selection. This mechanism allows the allocation of a greater survival to better individual: this is the survival-of-the-fittest mechanism we impose on our solution. There is several ways to select the chromosome which will be recombined. In our genetic algorithm, the roulette wheel selection will be used.

The principle of this method is that the better the chromosomes are, the more chances they have of being selected. Considering n chromosomes, each of them previously evaluated with a fitness value fi, we calculate the sum S of all the scores, as follows:

$$S = \sum_{i=0}^{n} f_i \qquad (2.25)$$

The probability Pi that a chromosome I will be selected is given simply by the chromosome fitness value divided by the sum S, as follows:

$$P_i = \frac{f_i}{S} \qquad (2.26)$$

Thus, the best chromosomes, e.g. with better fitness values, will be selected more frequently. A simple example with 5 chromosome illustrate the roulette wheel selection principle in table I, where the fitness value fi and the corresponding probability Pi of being selected are calculated.

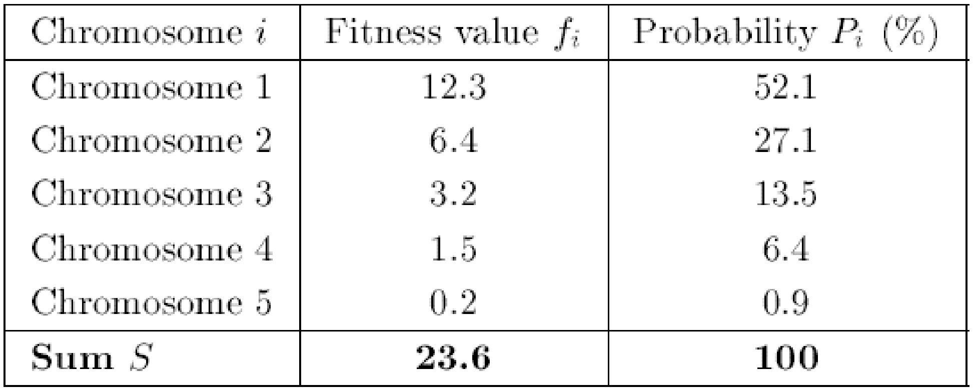
Crossover

The crossover is a process of taking genes from two parents, mixing them and producing an offspring. The simplest way is to choose randomly a crossover point. The child is produced by copying the first segment from the parent I and after the crossover point, the genes of the second parent are copied.

However, there are many others and complex ways to do crossover. The best is to have a lot of crossover points so as to have better chance to take the best from the both parents.

In our genetic algorithm, we decide to use the uniform-crossover, which use (n-I) crossover points for a chromosome containing n genes. For that a mask is created. Each bit of the mask is a random number between I or 2 and thus, determine from which parent each gene will be copied. (Figure 2.27)

Table 2.3 The fitness value *fi* of each chromosome i and the corresponding probability Pi of selection



Chromosome i	Fitness value f_i	Probability P_i (%)
Chromosome 1	12.3	52.1
Chromosome 2	6.4	27.1
Chromosome 3	3.2	13.5
Chromosome 4	1.5	6.4
Chromosome 5	0.2	0.9
Sum S	**23.6**	**100**

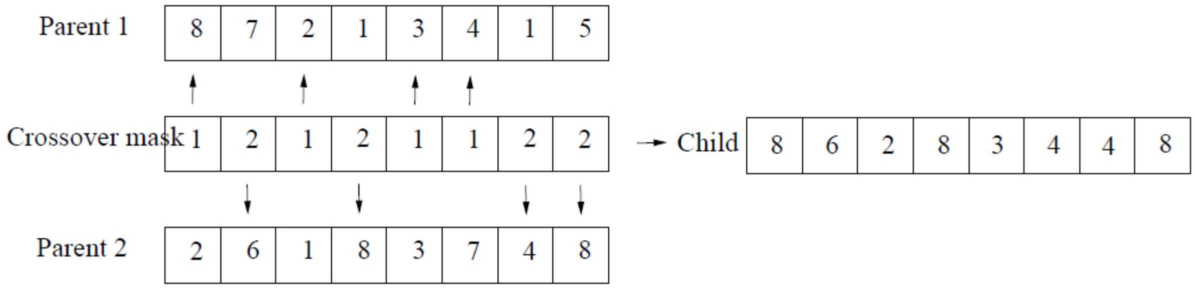


Figure 2.27 Principle of the Uniform crossover

Mutation

The mutation occurs after the crossover operation. It creates variants of few offsprings previously recombined and so introduce new genetic material. The probability of mutation must be low to stay in the neighborhood of the current solution. Otherwise, GA will not perform and better than a random search. In our study, we choose to mutate one offspring, selected randomly, in adding xi to one of its genes. The mutation is slight, between 0 to 0.201c so as to stay close to the current solution.

Population size

The population size is represent the number of chromosomes in a population. Of course, the more chromosomes there are, the more chances exist to have good solutions. However, the time for finding a solution has to be considered and if there are too many chromosomes, the genetic algorithm slows down. Sizes 20-30 are reported as best. Moreover, it has been shown that the best population size of chromosomes [26]. For these reasons, we choose a population size of 20 in the present study.